

The History of Physics  
Volume II  
*One Hundred of Selected Famous Papers*

---



The History of Physics  
Volume II  
*One Hundred of Selected Famous Papers*

---

**Javier Saramian**

Javier Saramian  
2014

**The History of Physics: One Hundred of Selected Famous Papers**

2014 by Javier Saramian

First Printing: 2014

1<sup>st</sup> Edition

Javier Saramian  
Gonzalo de Berceo 23-25  
Logroño, La Rioja 26005

To my lovely niece. I will try to plant the seed of the curiosity on you.

Also I would like to thank all the scientists, specially to the physicists, that tries to uncover the mysteries of the reality and the universe and to the divulgars that attempt to communicate all the beauty of this subject.

*Veritas In Hoc Scientia*



# Contents

---

## Volume I

<b>Book of Lemmas</b>	212	Archimedes BC	1
<b>On the Equilibrium of Planes. On Floating Bodies</b> <i>First mathematical physicist on record</i>			
<b>Discourse on the Method</b> <i>Framework for scientific method's guiding principles</i>	1637	Descartes, René	45
<b>Dialogues Concerning Two New Sciences</b> <i>Mechanics, kinematics, theory of inertia</i>	1638	Galilei, Galileo	59
<b>Philosophiae Naturalis Principia Mathematica</b> <i>Newton's laws of motion</i>	1687	Newton, Isaac	89
<b>Rules of Reasoning in Natural Philosophy</b> <i>Scientific method bases</i>	1729	Newton, Isaac	111
<b>On a General Method in Dynamics</b> <i>Development of the principle of "Varying Action"</i>	1834	Hamilton, William R.	113
<b>A Dynamical Theory of the Electromagnetic Field</b> <i>Theory of electromagnetism</i>	1865	Maxwell, James C.	179
<b>On the Relative Motion of the Earth and the Luminiferous Ether</b> <i>Evidence for an upper limit velocity of light</i>	1887	Michelson, Albert A. Morley, Edward W.	233
<b>On the Invisible Radiations Emitted by the Salts of Uranium</b> <i>Confirmation of the effect of spontaneous radioactivity</i>	1896	Becquerel, Henri	247
<b>Cathode Rays</b> <i>Discovery of the electron, the first elementary particle</i>	1897	Thomson, J. Joseph	253
<b>On an Improvement of Wien's Equation for the Spectrum</b> <i>Discovery of the Plank's radiation law for the energy spectrum. Beginnings of the quantum era in physics</i>	1900	Planck, Max	177
<b>On the Theory of the Energy Distribution Law of the Normal Spectrum</b> <i>Quantum hypothesis and explanation of the black body radiation spectrum</i>	1900	Planck, Max	281
<b>On the Law of Distribution of Energy in the Normal Spectrum</b> <i>Quantum hypothesis and final version of Planck's formula</i>	1901	Planck, Max	289
<b>Concerning an Heuristic Point of View Toward the Emission and Transformation of Light</b> <i>Explanation of the photoelectric effect using the light as corpuscular objects</i>	1905	Einstein, Albert	301
<b>On the Electrodynamics of Moving Bodies</b> <i>Invention of the theory of special relativity. Beginnings of the relativistic era in physics</i>	1905	Einstein, Albert	317
<b>Does the Inertia of a Body Depend on its Energy Content?</b> <i>Invention of the theory of special relativity, <math>E = mc^2</math></i>	1905	Einstein, Albert	343
<b>On the Relativity Principle and the Conclusions drawn from it</b> <i>Beginning of the long development of general relativity. Equivalence principle, gravitational redshift, and bending of light</i>	1907	Einstein, Albert	347
<b>On the Influence of Gravitation on the Propagation of Light</b> <i>General relativity replaces both special relativity and Newton's theory of gravitation. Equivalence principle only holds locally.</i>	1911	Einstein, Albert	361
<b>On an Expansion Apparatus for Making Visible the Tracks of Ionising Particles in Gases</b> <i>Invention of the cloud chamber to visualize tracks of ionizing particles</i>	1912	Wilson, Charles	371
<b>Penetrating Radiation in Seven Free Ballon Flights</b> <i>Conclusive evidence for the cosmic rays</i>	1912	Hess, Victor	393

<b>On the Constitution of Atoms and Molecules</b>	1913	Bohr, Niels	399
<i>Invention of the quantum theory of atomic spectra based on the Rutherford model of atomic structure - Bohr's atom</i>			
<b>On the Constitution of Atoms and Molecules 2</b>	1913	Bohr, Niels	423
<i>Bohr's quantum theory of atomic spectra. Evidence that radioactivity is a nuclear property</i>			
<b>The Structure of the Atom</b>	1914	Rutherford, Ernest	449
<i>Confirmation of the existence of atomic nuclei. First indication of the existence of the proton</i>			
<b>On the gravitational field of a mass point according to Einstein's theory</b>	1916	Schwarzschild, Karl	461
<i>Schwarzschild metric</i>			
<b>The Foundation of the General Theory of Relativity</b>	1916	Einstein, Albert	469
<i>Theory of general relativity</i>			
<b>Invariant Variant Problems</b>	1918	Noether, Emmy	525
<i>Noether theorem connecting symmetries and conserved quantities</i>			
<b>On the Quantum Theory of Line Spectra</b>	1918	Bohr, Niels	539
<i>Bohr's invention of correspondence principle</i>			
<b>Collision of a Particle with Light Atoms. IV. Anomalous Effect in Nitrogen</b>	1919	Rutherford, Ernest	579
<i>Discovery of the proton. Evidence for it as a nucleus constituent</i>			
<b>Nuclear Constitution of Atoms</b>	1920	Rutherford, Ernest	587
<i>Rutherford neutron hypothesis</i>			
<b>On the Curvature of Space</b>	1922	Friedmann, Aleksandr	617
<i>Friedmann Cosmology</i>			
<b>Waves and Quanta</b>	1923	De Broglie, Louis	623
<i>Suggestion of the corpuscular-wave dualism for electrons</i>			
<b>The Spectrum of Scattered X-Rays</b>	1923	Compton, Arthur	627
<i>Direct experimental confirmation that the photon is an elementary particle, the Compton effect</i>			
<b>Investigations on X Rays and Beta Rays by the Cloud Method. Part 1. - X Rays</b>	1923	Wilson, Charles	633
<i>Experimental confirmation of the ionization process predicted by Compton for a corpuscular photon</i>			
<b>Planck's Law and Light Quantum Hypothesis</b>	1924	Bose, Satyendra	657
<i>Discovery of Bose-Einstein quantum statistics for particles of integer spins. New derivation of Planck's radiation law</i>			
<b>On the Connexion between the Completion of Electron Groups in an Atom with the Complex Structure of Spectra</b>	1925	Pauli, Wolfgang	661
<i>Discovery of the exclusion principle - the Pauli principle</i>			
<b>Quantum-Theoretical Re-Interpretation of Kinematic and Mechanical Relations</b>	1925	Heisenberg, Werner	675
<i>Foundation of quantum mechanics, Heisenberg approach</i>			
<b>On Quantum Mechanics</b>	1925	Bohr, Niels	691
<i>Invention of matrix formalism for the Heisenberg quantum mechanics. Systems with one degree of freedom</i>			
<b>On Quantum Mechanics II</b>	1925	Bohr, Niels	721
<i>Matrix formalism with arbitrary many degrees of freedom</i>			

## Volume II

<b>On Quantizing an Ideal Monatomic Gas</b>	1926	Fermi, Enrico	787
<i>Invention of statistics for ensembles of particles obeying Pauli principle - Fermi-Dirac quantum statistics</i>			
<b>On the Quantum Mechanics of Collisions</b>	1926	Bohr, Niels	797
<i>Statistical interpretation, probability density of quantum mechanics. Quantum theory of scattering. Born approximation</i>			
<b>Quantisation as a Problem of Proper Values - Part 1</b>	1926	Schrödinger, Erwin	801
<i>Creation of wave mechanics. Invention of the Schrödinger wave equation</i>			
<b>The Physical Content of Quantum Kinematics and Mechanics</b>	1927	Heisenberg, Werner	813
<i>Heisenberg discovery of the uncertainty principle</i>			
<b>The Quantum Theory of Dispersion</b>	1927	Dirac, Paul	837
<i>Foundations of quantum electrodynamics - QED</i>			
<b>The Quantum Theory of the Electron</b>	1928	Dirac, Paul	857
<i>Discovery of the relativistic wave equation for the electron. Prediction of the magnetic moment of the electron</i>			
<b>A Relation between Distance and Radial Velocity</b>	1929	Hubble, Edwin	873
<i>Evidence for the expansion of the universe</i>			
<b>Gravitation and the Electron</b>	1929	Weyl, Hermann	879
<i>Combination of electrodynamics and gravitation</i>			
<b>The Production of High Speed Protons Without the use of High Voltages</b>	1931	Lawrence, Ernest	893
<i>Tests of the first cyclotron</i>			
<b>Quantised Singularities in the Electromagnetic Field</b>	1931	Dirac, Paul	895
<i>Prediction of the anti-electron (<math>e^+</math>), anti-proton (anti-<math>p</math>), and an indication of the possible existence of magnetic monopoles</i>			
<b>A 1,500,000 Volt Electrostatic Generator</b>	1931	Van de Graaff, Robert	909
<i>Invention of the Van de Graaff electrostatic accelerator</i>			
<b>The Existence of a Neutron</b>	1932	Chadwick, James	911
<i>Discovery of the neutron</i>			
<b>The Electrostatic Production of High Voltage for Nuclear Investigations</b>	1933	Van de Graaff, Robert Compton, Arthur	927
<i>Invention of electrostatic accelerators. Further development</i>			
<b>The Positive Electron</b>	1933	Anderson, Carl	943
<i>Discovery of the positron, the first antiparticle, predicted by Dirac</i>			
<b>The Highly Collapsed Configurations of a Stellar Mass</b>	1935	Chandrasekhar, Subrahmanyan	951
<i>Chandrasekhar Limit</i>			
<b>Can Quantum Mechanical Description of Physical Reality be considered complete?</b>	1935	Einstein, Albert Podolsky, Boris Rosen, Nathan	971
<i>EPR paradox</i>			
<b>On the Interaction of Elementary Particles</b>	1935	Yukawa, Hideki	975
<i>Yukawa field theory of nuclear forces. Prediction of heavy quanta, the pion particles, as mediators of strong interactions</i>			
<b>Visible Radiation Produced by Electrons Moving in a Medium with Velocities Exceeding that of Light</b>	1937	Cherenkov, Pavel	985
<i>Confirmation of the Frank-Tamm theory of the Vavilov- Cerenkov effect</i>			
<b>New Evidence for the Existence of a Particle of Mass</b>	1937	Street, Jabez Stevenson, E	989
<i>Muon existence confirmation</i>			
<b>On Massive neutron cores</b>	1939	Oppenheimer, Robert	995
<i>Prediction of neutron stars</i>			
<b>Forces in Molecules</b>	1939	Feynman, Richard	1003
<i>Feynman-Hellmann theorem, as an efficient approach to the calculation of forces in molecules</i>			

<b>On Continued Gravitational Contraction</b>	1939	Oppenheimer, Robert	1007
<i>First notion of black hole</i>			
<b>Acceleration of Electrons by Magnetic Induction</b>	1940	Kerst, Donald	1013
<i>Kerst proposal for betatron accelerator</i>			
<b>The Connection Between Spin and Statistics</b>	1940	Pauli, Wolfgang	1017
<i>Theorem on the connection between spin and statistics</i>			
<b>Expanding Universe and the Origin of Elements</b>	1946	Gamov, George	1031
<i>Explanation to the observed chemical elements abundance-curve by unequilibrium process of elements formation. Birth of the Big Bang model</i>			
<b>Space-Time Approach to Non-Relativistic Quantum Mechanism</b>	1948	Feynman, Richard	1035
<i>Path integral formalism</i>			
<b>A Generalized Theory of Gravitation</b>	1948	Einstein, Albert	1071
<i>Summary of the General Relativity</i>			
<b>The Origin of Chemical Elements</b>	1948	Gamov, George	1077
<i>Theory of big-bang nucleosynthesis</i>			
<b>A Relativistic Cut-Off for Quantum Electrodynamics</b>	1948	Feynman, Richard	1079
<i>Proposal to modify classical electrodynamics to a form suitable for quantization</i>			
<b>Space-Time Approach to Quantum Electrodynamics</b>	1949	Feynman, Richard	1089
<i>Development of the covariant quantum electrodynamic theory. Feynman method</i>			
<b>The Theory of Positrons</b>	1949	Feynman, Richard	1143
<i>Creation of the covariant quantum electrodynamic theory. Feynman method</i>			
<b>Mathematical Formulation of the Quantum Theory of Electromagnetic Interaction</b>	1950	Feynman, Richard	1155
<i>Mathematical proof of the validity of the Feynman rules for calculations of amplitudes in QED</i>			
<b>Equation of State Calculations by Fast Computing Machines</b>	1953	Teller, Edward	1173
<i>The birth of computational physics</i>			
<b>Conservation of Isotopic Spin and Isotopic Gauge Invariance</b>	1954	Yang, Chen Ning Mills, Robert	1179
<i>Introduction of local gauge isotopic spin invariance in quantum field theory: Yang-Mills theory</i>			
<b>Relative State Formulation of Quantum Mechanics</b>	1957	Everett, Hugh	1185
<i>"Many worlds" interpretation</i>			
<b>Dynamical Model of Elementary Particles Based on an Analogy with Superconductivity. I</b>	1961	Nambu, Yoichiro	1207
<i>Nambu-Jona-Lasinio nonlinear model of hadrons</i>			
<b>Dynamical Model of Elementary Particles Based on an Analogy with Superconductivity. II</b>	1961	Nambu, Yoichiro	1221
<i>Nambu-Jona-Lasinio nonlinear model of hadrons</i>			
<b>A Schematic Model of Baryons and Mesons</b>	1964	Gell-Man Murray	1231
<i>Introduction of quarks as fundamental building blocks for hadrons.</i>			
<b>Broken Symmetries and the Masses of Gauge Bosons</b>	1964	Higgs, Peter	1233
<i>Higgs mechanism of mass generation for vector gauge fields</i>			
<b>Inhomogeneous Electron</b>	1964	Hohenberg, Pierre Walter, Kohn	1235
<i>Density Functional Theory</i>			
<b>On the Einstein Podolsky Rosen Paradox</b>	1964	Bell, John	1243
<i>Bell theorem</i>			
<b>Three Triplet Model with Double SU(3) Symmetry</b>	1965	Nambu, Yoichiro	1249
<i>Suggestion of the existence of three triplets of quarks</i>			
<b>Self-Consistent Equations including Exchange and Correlation Effects</b>	1965	Kohn, Walter Sham, Lu Jeu	1255
<i>Density Functional Theory</i>			
<b>Quantum Theory of Gravity. I. The Canonical Theory</b>	1967	deWitt, Bryce	1261
<i>Wheeler-DeWitt equation of quantum gravity</i>			

<b>A Model for Leptons</b>	1967	Weinberg, Steven	1297
<i>Lagrangian for the electroweak synthesis, estimations of the W and Z masses</i>			
<b>Construction of a Crossing-symmetric, Regge-behaved Amplitude for Linearly Rising Trajectories</b>	1968	Veneziano, Gabriele	1301
<i>First formulation of string theory</i>			
<b>On the interpretation of measurement in quantum theory</b>	1969	Zeh, Heinz-Dieter	1311
<i>Quantum decoherence</i>			
<b>Renormalization of Massless Yang-Mills Fields</b>	1971	Hooft, Gerard 't	1319
<i>Rigorous proofs of renormalizability of the massless Yang-Mills quantum gauge fields theory</i>			
<b>Black Holes and Entropy</b>	1972	Bekenstein, Jacob	1347
<i>Black hole thermodynamics</i>			
<b>Ultraviolet Behavior of Non Abelian Gauge Theory</b>	1973	Gross, David	1361
<i>Asymptotic freedom</i>			
<b>Unity of All Elementary-Particle Forces</b>	1974	Georgi, Howard	1365
<i>Grand unified theory of all forces except gravity</i>		Glashow, Sheldon	
<b>Particle Creation by Black Holes</b>	1975	Hawking, Stephen	1369
<i>Hawking radiation and black holes temperature</i>			
<b>Inflationary universe. A possible solution to the horizon and flatness problems</b>	1980	Guth, Alan	1391
<i>Theory of inflation</i>			
<b>Simulating Physics with Computers</b>	1981	Feynman, Richard	1401
<i>Argument for quantum computation</i>			
<b>A new inflationary universe scenario, A possible solution of the horizon, flatness, homogeneity, isotropy and primordial monopole problems</b>	1981	Linde, Andrei	1423
<i>Extended theory of inflation</i>			
<b>Quantum theory, the Church-Turing principle and the universal quantum computer</b>	1985	Deutsch, David	1429
<i>Foundation of the universal quantum Turing model of computation</i>			
<b>New variables for classical and quantum gravity</b>	1986	Ashtekar, Abhay	1449
<i>Ashtekar variables</i>			
<b>Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels</b>	1992	Bennett, Charles	1453
<i>Brassard, Gilles</i>			
<i>Superdense coding</i>			
<b>Dimensional Reduction in Quantum Gravity</b>	1993	Hooft, Gerard 't	1459
<i>Holographic principle</i>			
<b>String Theory Dynamics in Various Dimensions</b>	1995	Witten, Edward	1473
<i>M-theory</i>			
<b>The Large N Limit of Superconformal field</b>	1997	Maldacena, Juan	1529
<i>AdS/CFT Correspondence</i>			
<b>The Hierarchy Problem and New Dimensions at a Millimeter</b>	1998	Arkani-Hammed, Nima	1551
<i>Proposed an alternative scenario to explain the weakness of gravity relative to the other forces</i>			



## On Quantizing an Ideal Monatomic Gas

E. Fermi

(Received 1926)

In classical thermodynamics the molecular heat (an constant volume) is

$$c = \left(\frac{3}{2}\right)k. \quad (1)$$

If, however, we are to apply Nernst's heat theorem to a gas we must consider (1) merely as an approximation for high temperatures since  $c$  must vanish in the limit as  $T = 0$ . We are therefore forced to assume that the motion of a molecule in an ideal gas is quantized; this quantization manifests itself for low temperatures by certain degeneracy phenomena so that the specific heat and the equation of state depart from their classical counterparts.

The aim of the present paper is to present a method of quantization of an ideal gas which, according to our opinion, is as independent of arbitrary assumptions about the statistical of the gas molecules as is possible.

In recent times, numerous attempts have been made to determine the equation of state of a perfect gas. The equations of state of the various authors and ours differ from each other and from the classical equation of state

$$PV = NkT$$

by the terms, which become appreciable only at very low temperatures and high pressures; unfortunately, real gases depart most strongly from ideal gases under these conditions so that the significant degeneracy phenomena have not been observable up until now. In any case, it may well be that a deeper knowledge of the equation of state may enable us to separate the degeneracy from the remaining deviations from the equation  $PV = NkT$  so that it may be possible to decide experimentally which of the degeneracy theories is correct.

To apply the quantum rules to the motions of the molecules, we can proceed in various ways; the result, however, is always the same. For example, we may picture the molecules as being enclosed in a parallelopiped container with elastically reflecting walls; then the motion of the molecules flying back and forth between the walls is conditionally periodic and can therefore be quantized; more generally, we may picture the molecules as moving in an external force field, such that their motion is conditionally periodic; the assumption that the gas is ideal permits us to neglect the interactions of the molecules, so that their mechanical motions occur only under the influence of the external field. It is clear, however, that the quantization of the molecular motion made under the assumption of the complete independence of the molecules from one another is not sufficient to account for the expected degeneracy. We can see this best in the example of molecules in a container if we note that as linear dimensions of the container increase, the energy levels of the quantum states of each molecule become denser and denser, so that for vessels of macroscopic dimensions all influences of the discontinuity of the energy values practically disappear. This influence, moreover, depends on the volume of the container, even if the number of molecules in it are so chosen that the density remains constant.

By analyzing this state of affairs quantitatively, we can convince ourselves that we only then obtain a degeneracy of the expected magnitude when we choose the vessel so small that it contains, on the average, just one molecule.

We therefore surmise that the quantization of ideal gases requires an addition to the Sommerfeld quantum condition.

Now recently Pauli, following upon an investigation by E.C. Stoner, proposed the rule that if an electron inside an atom has quantum numbers (including the magnetic quantum number) with definite values, then no other electron can exist in the atom in an orbit which is characterized by the same quantum numbers. In other words, a quantum state (in an external magnetic field) is already completely filled by a single electron.

Since this Pauli rule has proved extremely fruitful in the interpretation of spectroscopic phenomena, we want to see whether it may not also be useful in the problem of the quantization of ideal gases.

We shall show that this is, indeed, the case, and that the application of Pauli's rule allows us to present a completely consistent theory of the degeneracy of an ideal gas.

We therefore assume in the following that, at most, one molecule with given quantum numbers can exist in our gas: as quantum numbers we must take into account not only those that determine the internal motions of the molecule but also the numbers that determine its translational motion.

We must first place our molecules in a suitable external force field so that their motion is conditionally periodic. This can be done in an infinitude of ways; since, however, the result does not depend on the choice of the force field, we shall impose on the molecules a central elastic force directed toward a fixed point  $O$  (the coordinate origin) so that each molecule becomes a harmonic oscillator. This central force will keep our gas mass in the neighborhood of  $O$ ; the gas density will decrease with increasing distance from  $O$  and vanish at infinity. If

$\nu$  is proper frequency of the oscillators, then the force exerted on the molecules is

$$4\pi^2\nu^2mr$$

where  $m$  is the mass of the molecule and  $r$  its distance from  $O$ . The potential energy of the attractive force is then

$$u = 2\pi^2\nu^2mr^2$$

Let  $s_1, s_2, s_3$  be the quantum numbers of a molecule oscillator. These quantum numbers are essentially not sufficient to characterize the molecule, for we must add to these the quantum numbers of the internal motions. We limit ourselves, however, to monatomic molecules and assume, in addition, that all the molecules in our gas are in the ground state and that this state is single (does not split in a magnetic field). We need not worry about the internal motion then, and we may then consider our molecules simply as mass points. The Pauli rule, therefore, states in our case that in the entire mass of gas at most only one molecule can have the given quantum numbers  $s_1, s_2, s_3$ .

The total energy of this molecule is given by

$$w = h\nu(s_1 + s_2 + s_3) = h\nu s. \quad (2)$$

The total energy can thus be an arbitrary integral multiple of  $h\nu$ ; the value  $sh\nu$ , however, can be realized in many ways. Each realization implies a solution of the equation

$$s = s_1 + s_2 + s_3 \quad (3)$$

where  $s_1, s_2, s_3$  can assume the values 1, 2, 3 . . . We know that (3) has

$$Q_s = \frac{(s+1)(s+2)}{2} \quad (4)$$

solutions. The energy 0 can thus be realized in one way, the energy  $h\nu$  in three ways, the energy  $2h\nu$  in six ways, and so on. We shall simply call a molecule with energy  $sh\nu$  an  $((s))$ -molecule.

According to our assumption, there can be in our entire gas mass only  $Q_s((s))$ -molecules; thus, at most, one molecule with energy zero, at most, three with energy  $h\nu$  at most, six with energy  $2h\nu$ , and so on.

To see early the results of this state of affairs, we consider the extreme case in which the absolute temperature of our gas is zero. Let  $N$  be the number of molecules. At absolute zero our gas must be in its lowest energy state. If there were no restrictions on the number of molecules of a given energy, then every molecule would be in a state of zero energy ( $s_1 = s_2 = s_3 = 0$ ). According to the foregoing, however, at most, only one molecule can have zero energy; hence, if  $N$  were 1, then this single molecule would have energy zero; if  $N$  were 4, one molecule would have energy zero and the three other would occupy the three available places with energy  $h\nu$ ; if  $N$  were 10 one molecule would be in the zero energy position, three others in the three places with energy  $h\nu$ , and the six remaining ones in the six places with energy  $2h\nu$  and so on.

At the absolute zero point, our gas molecules arrange themselves in a kind of shell-like structure which has a certain analogy to the shell-like arrangement of electrons in an atom with many electrons.

We now want to investigate how a certain amount of energy

$$W = Eh\nu \quad (5)$$

( $E$  = integer) is distributed among our molecules.

Let  $N_s$  be the number of molecules in a state with energy  $sh\nu$ . According to our assumption

$$N_s \leq Q_s \quad (6)$$

We have, further, the equations

$$\sum N_s = N \quad (7)$$

$$\sum sN_s = E \quad (8)$$

which state that the total number and total energy of the molecules are  $N$  and  $Eh\nu$ , respectively.

We now want to calculate the number  $P$  of arrangements of our  $N$  molecules for which  $N_0$  are at places with energy 0,  $N_1$  at places with energy  $h\nu$ ,  $N_2$  at places with energy  $2h\nu$ , etc. Two such arrangements are to be considered identical if the places occupied by the molecules are the same; thus two arrangements which differ only in a permutation among the molecules in their places are to be considered as one. If we considered two such arrangements as different, we would have to multiply  $P$  by the constant  $N!$ ; we can easily see, however, that this can have no influence on what follows. In the above-defined sense, the number of arrangements of  $N_s$  molecules among the  $Q_s$  places of energy,  $sh\nu$  is given by

$$\binom{Q_s}{N_s}.$$

We therefore find for  $P$  the expression

$$P = \binom{Q_s}{N_0} \binom{Q_1}{N_1} \binom{Q_2}{N_2} \dots = \Pi \binom{Q_s}{N_s}. \quad (9)$$

We obtain the most probable values of the  $N_s$  by seeking the maximum of  $P$  under the constraints (7) and (8). By applying Stirling's theorem we may write sufficient approximation for our case

$$\log P = \sum \log \binom{Q_s}{N_s} = - \sum \left( N_s \log \frac{N_s}{Q_s - N_s} + Q_s \log \frac{Q_s - N_s}{Q_s} \right) \quad (10)$$

We thus seek the values of  $N_s$  that satisfy (7) and (8) and for which  $\log P$  becomes a maximum. We find

$$\alpha e^{-\beta s} = \frac{N_s}{Q_s - N_s}$$

where  $\alpha$  and  $\beta$  are constants. This equation gives us

$$N_s = Q_s \cdot \frac{\alpha e^{-\beta s}}{1 + \alpha e^{-\beta s}} \quad (11)$$

The values of  $\alpha$  and  $\beta$  can be found from equation (7) and (8) or, conversely, we may consider  $\alpha$  and  $\beta$  as given; then (7) and (8) determine the total number and total energy of our configuration. We find, namely,

$$N = \sum_0^{\infty} Q_s \frac{\alpha e^{-\beta s}}{1 + \alpha e^{-\beta s}} \quad (12)$$

$$\frac{W}{h\nu} = E = \sum_0^{\infty} s \cdot Q_s \frac{\alpha e^{-\beta s}}{1 + \alpha e^{-\beta s}}$$

The absolute temperature  $T$  of the gas is a function of  $N$  and  $E$  or also of  $\alpha$  and  $\beta$ . This function can be determined by two methods, which, however, lead to the same result. We could, for example, according to the Boltzmann entropy principle set

$$S = k \log P$$

and then calculate the temperature from the formula

$$T = \frac{dW}{dS}$$

This method, however, has the disadvantage common to all methods based on the Boltzmann principle, that for its application we must make a more or less arbitrary assumption about the probability of a state. Therefore, we proceed as follows: we note that the density of our gas is a function of the distance which vanishes for infinite distances. For infinitely large  $r$ , therefore, the degeneracy phenomena also vanish and the statistics of our gas go over to classical statistics. In particular, for  $r = \infty$  the mean kinetic energy of a molecule must become  $(2/3)kT$  and the velocity distribution must go over to the Maxwellian. We can thus obtain the temperature from the distribution of velocities in the region of infinitesimal densities; and since the entire gas is at the same constant temperature, we then at the same time obtain the temperature of the high density region also. For this determination we use, so to speak, a gas thermometer with an infinitely attenuated ideal gas.

To begin with, we calculate the density of molecules with kinetic energy between  $L$  and  $L + dL$  at the distance  $r$ . The total energy of these molecules lies, according to (1), between

$$L + 2\pi^2\nu^2mr^2 \quad \text{and} \quad L + 2\pi^2\nu^2mr^2 + dL.$$

Now the total energy of a molecule is  $sh\nu$ . For our molecules  $s$  must therefore lie between  $s$  and  $s + ds$ , where

$$s = \frac{L}{h\nu} + \frac{2\pi^2\nu^2mr^2}{h} \quad ds = \frac{dL}{h\nu} \quad (13)$$

We now consider a molecule whose motion is characteristic by the quantum numbers  $s_1, s_2, s_3$ . Its coordinates  $x, y, z$  are given by

$$x = \sqrt{Hs_1} \cos(2\pi\nu t - \alpha_1), \quad y = \sqrt{Hs_2} \cos(2\pi\nu t - \alpha_2) \quad (14)$$

$$z = \sqrt{Hs_3} \cos(2\pi\nu t - \alpha_3)$$

as functions of the time. Here

$$H = \frac{\hbar}{2\pi^2\nu m}; \quad (15)$$

$\alpha_1, \alpha_2, \alpha_3$  are phase constants which may take on all sets of values with equal probability. From this and from equation (14) it follows that

$$|x| \leq \sqrt{Hs_1}, |y| \leq \sqrt{Hs_2}, |z| \leq \sqrt{Hs_3},$$

and that the probability that  $x, y, z$  lie between the limits  $x$  and  $x + dx$ ,  $y$  and  $y + dy$ ,  $z$  and  $z + dz$ , has the value

$$\frac{dxdydz}{\pi^3 \sqrt{(Hs_1 - x^2)(Hs_2 - y^2)(Hs_3 - z^2)}}$$

If we do not know the individual values of  $s_1, s_2, s_3$  but only their sum, then our probability is given by

$$\frac{1}{Q_s} \cdot \frac{dxdydz}{\pi^3} \cdot \sum \frac{1}{\sqrt{(Hs_1 - x^2)(Hs_2 - y^2)(Hs_3 - z^2)}} \quad (16)$$

The sum is to be extended over all integer solutions of equation (3) which satisfy the inequalities

$$Hs_1 \geq x^2, \quad Hs_2 \geq y^2, \quad Hs_3 \geq z^2$$

If we multiply the probability (16) with the number  $N_s$  of  $((s))$  – molecules, we obtain the number of  $((s))$  – molecules in the volume element  $dxdydz$ . Taking account of (11) we thus find that the density of  $((s))$  – molecules at the position  $x, y, z$  is given by

$$N_s = \frac{\alpha e^{-\beta s}}{1 + \alpha e^{-\beta s}} \cdot \frac{1}{\pi^3} \cdot \sum \frac{1}{\sqrt{(Hs_1 - x^2)(Hs_2 - y^2)(Hs_3 - z^2)}}$$

For sufficiently large  $s$  we can replace the sum by a double integral; after carrying out the integration we find

$$N_s = \frac{2}{\pi^2 H^2} \cdot \frac{\alpha e^{-\beta s}}{1 + \alpha e^{-\beta s}} \cdot \sqrt{Hs - r^2}.$$

Using (13) and (15) we now find that the density of molecules with kinetic energy between  $L$  and  $L + dL$  at the position  $x, y, z$  is given by the following expression

$$N(L)dL = N_s ds = \frac{2\pi(2m)^{3/2}}{h^3} \cdot \sqrt{L} dL \cdot \frac{\alpha e^{-\frac{2\pi^2\nu m \beta r^2}{h}} e^{-\frac{\beta L}{h\nu}}}{1 + \alpha e^{-\frac{2\pi^2\nu m \beta r^2}{h}} e^{-\frac{\beta L}{h\nu}}} \quad (17)$$

This formula must be compared with the classical expression for the Maxwellian distribution:

$$N^*(L)dL = K\sqrt{L} dLe^{-L/kT} \quad (17')$$

We see then that in the limit for  $\nu = \infty$  (17) goes over into (17') if we just set

$$\beta = \frac{h\nu}{kT} \quad (18)$$

Now (17) can be written as follows:

$$N(L)dL = \frac{(2\pi)(2m)^{3/2}}{h^3} \cdot \sqrt{L} dL \cdot \frac{Ae^{-L/kT}}{1 + Ae^{-L/kT}} \quad (19)$$

where

$$A = \alpha e^{-\frac{2\pi^2\nu^2mr^2}{kT}} \quad (20)$$

The total density of molecules at the distance  $r$  now becomes

$$N = \int_0^\infty N(L) dL = \frac{(2\pi mkT)^{3/2}}{h^3} F(A), \quad (21)$$

where we have placed

$$F(A) = \frac{2}{\sqrt{\pi}} \cdot \int_0^\infty \frac{A\sqrt{x}e^{-x}dx}{1 + Ae^{-x}} \quad (22)$$

The mean kinetic energy of the molecules at the distance  $r$  is

$$\bar{L} = \frac{1}{N} \int_0^\infty LN(L)dL = \left(\frac{3}{2}\right) \cdot kT \cdot \frac{G(A)}{F(A)} \quad (23)$$

where

$$G(A) = \frac{4}{3\sqrt{\pi}} \int_0^\infty \frac{Ax^{3/2}e^{-x}dx}{1 + Ae^{-x}}. \quad (24)$$

Through (21) we can determine  $A$  as a function of density and temperature; when we put this into (19) and (20) we obtain the velocity distribution and the mean kinetic energy as a function of density and temperature.

To obtain the equation of state we use the virial theorem. According to this pressure is given by

$$p = \frac{2}{3} \cdot N\bar{L} = NkT \cdot \frac{G(A)}{F(A)}; \quad (25)$$

again  $A$  is to be found from (12) as a function of density and temperature.

Before we go further we give some of the mathematical properties of  $F(A)$  and  $G(A)$ .

For  $A \leq 1$  we can express both functions by convergent series

$$\begin{cases} F(A) = A - \frac{A^2}{2^{3/2}} + \frac{A^3}{3^{3/2}} - \dots \\ G(A) = A - \frac{A^2}{2^{3/2}} + \frac{A^3}{3^{3/2}} - \dots \end{cases} \quad (26)$$

For large  $A$  we have the asymptotic expressions

$$\begin{cases} F(A) = \frac{4}{3\sqrt{\pi}} (\log A)^{3/2} \left[ 1 + \frac{\pi^2}{8(\log A)^2} + \dots \right], \\ G(A) = \frac{8}{15\sqrt{\pi}} (\log A)^{5/2} \left[ 1 + \frac{5\pi^2}{8(\log A)^2} + \dots \right]. \end{cases} \quad (27)$$

Further, the relationship

$$\frac{dG(A)}{F(A)} = d \log A \quad (28)$$

holds.

We must still introduce another function  $P(\Theta)$  defined by

$$P(\Theta) = \Theta \cdot \frac{G(A)}{F(A)}, \quad F(A) = \frac{1}{\Theta^{3/2}} \quad (29)$$

For very large and very small  $\theta$  respectively,  $P(\theta)$  can be calculated from the approximations

$$\begin{cases} P(\Theta) = \Theta \left\{ 1 + \frac{1}{2^{5/2}\Theta^{3/2}} + \dots \right\} \\ P(\Theta) = \frac{3^{3/2}\pi^{1/3}}{5 \cdot 2^{1/3}} \frac{3^{2/3}\pi^{1/3}}{5 \cdot 2^{1/3}} \left\{ 1 + \frac{5 \cdot 2^{2/3}\pi^{4/3}}{3^{7/3}} \Theta^2 + \dots \right\} \end{cases} \quad (30)$$

Using (29), (28), (27), we see further that

$$\int_0^\Theta \frac{dP(\Theta)}{\Theta} = \frac{1}{3} \cdot \frac{G(A)}{F(A)} - \frac{2}{3} \log A. \quad (31)$$

We can now eliminate  $A$  from the equation of state (25) and (23) and we obtain the pressure and the mean kinetic energy as explicit functions of density and temperature:

$$p = \frac{h^2 N^{5/3}}{2\pi m} \cdot P \cdot \left( \frac{2\pi m k T}{h^2 N^{2/3}} \right) \quad (32)$$

$$\bar{L} = \frac{h^2 N^{2/3}}{2\pi m} \cdot P \cdot \left( \frac{2\pi m k T}{h^2 N^{2/3}} \right) \quad (33)$$

In the limit of weak degeneracy ( $T$  large and  $N$  small) the equation of state has the following form:

$$p = N k T \left\{ 1 + \left( \frac{1}{16} \right) \cdot \frac{h^3 N}{(\pi m k T)^{3/2}} + \dots \right\}. \quad (34)$$

The pressure is thus larger than the classical pressure  $P = (NkT)$ . For an ideal gas with the atomic weight of helium at  $T = 5^\circ$  and a pressure of 10 atm, the difference is about 15%.

In the limit of large degeneracy, (32) and (33) become

$$\begin{aligned} p &= \left(\frac{1}{20}\right)^{2/3} \cdot \frac{h^2 N^{5/3}}{m} + \frac{2^{4/3}}{3^{5/2}} \pi^{8/3} \cdot \frac{m N^{1/3} k^2 T^2}{h^2} + \dots \\ \bar{L} &= \left(\frac{3}{40}\right)^{2/3} \cdot \frac{h^2 N^{2/3}}{m} + \frac{2^{1/3}}{3^{2/3}} \pi^{8/3} \cdot \frac{m k^2 T^3}{h^2 N^{2/3}} + \dots \end{aligned} \quad (35)$$

From this we see that the degeneracy leads to a zero point pressure and a zero point energy.

From (35) we can also obtain the specific heat at low temperatures.

We find

$$C_v = \frac{d\bar{L}}{dT} = \frac{2^{4/3} \pi^{8/3}}{3^{2/3}} \frac{m k^2 T}{h^2 N^{2/3}} + \dots \quad (36)$$

The specific heat vanishes at absolute zero and is proportional to the absolute temperature at low temperatures . . .



## I.2 ON THE QUANTUM MECHANICS OF COLLISIONS

[Preliminary communication]<sup>†</sup>

MAX BORN

Through the investigation of collisions it is argued that quantum mechanics in the Schrödinger form allows one to describe not only stationary states but also quantum jumps.

Heisenberg's quantum mechanics has so far been applied exclusively to the calculation of stationary states and vibration amplitudes associated with transitions (I purposely avoid the word "transition probabilities"). In this connection the formalism, further developed in the meantime, seems to be well validated. However, questions of this kind deal with only one aspect of quantum theory. Beside them there shows up as equally important the question of the nature of the "transitions" themselves. On this point opinions seem to be divided. Many assume that the problem of transitions is not encompassed by quantum mechanics in its present form, but that here new conceptual developments will be necessary. I myself, impressed with the closed character of the logical nature of quantum mechanics, came to the presumption that this theory is complete and that the problem of transitions must be contained in it. I believe that I have now succeeded in proving this.

Bohr has already directed attention to the fact that all difficulties of principle associated with the quantum approach which meet us in the emission and absorption of light by atoms also occur in the interaction of atoms at short distances and consequently in collision processes. In collisions one deals not with mysterious wave fields, but exclusively with systems of material particles, subject to the formalism of quantum mechanics. I therefore attack the problem of investigating more closely the interaction of the free particle ( $\alpha$ -ray or electron) and an arbitrary atom and of determining whether a description of a collision is not possible within the framework of existing theory.

Of the different forms of the theory only Schrödinger's has proved suitable for this process, and exactly for this reason I might regard it as the deepest formulation of the quantum laws. The course of my reasoning is the following.

If one wishes to calculate quantum mechanically the interaction of two systems,

<sup>†</sup> This report was originally intended for *die Naturwissenschaften*, but could not be accepted there for lack of space. I hope that its publication in this journal [*Zeitschrift für Physik*] does not seem out of place [M.B.].

Originally published under the title, "Zur Quantenmechanik der Stossvorgänge," *Zeitschrift für Physik*, 37, 863–67 (1926); reprinted in *Dokumente der Naturwissenschaft*, 1, 48–52 (1962) and in M. Born (1963); translation into English by J.A.W. and W.H.Z., 1981.

then, as is well known, one cannot, as in classical mechanics, pick out a state of the one system and determine how this is influenced by a state of the other system, since all states of both systems are coupled in a complicated way. This is true also in an aperiodic process, such as a collision, where a particle, let us say an electron, comes in from infinity and then goes off to infinity. There is no escape from the conclusion that, as well before as after collision, when the electron is far enough away and the coupling is small enough, a definite state must be specifiable for the atom and likewise a definite rectilinear motion for the electron. The problem is to formulate mathematically this asymptotic behavior of the coupled particles. I did not succeed in doing this with the matrix form of quantum mechanics, but did with the Schrödinger formulation.

According to Schrödinger, the atom in its  $n$ th quantum state is a vibration of a state function of fixed frequency  $W_n^0/h$  spread over all of space. In particular, an electron moving in a straight line is such a vibratory phenomenon which corresponds to a plane wave. When two such waves interact, a complicated vibration arises. However, one sees immediately that one can determine it through its asymptotic behavior at infinity. Indeed one has nothing more than a "diffraction problem" in which an incoming plane wave is refracted or scattered at an atom. In place of the boundary conditions which one uses in optics for the description of the diffraction diaphragm, one has here the potential energy of interaction between the atom and the electron.

The task is clear. We have to solve the Schrödinger wave equation for the system atom-plus-electron subject to the boundary condition that the solution in a preselected direction of electron space goes over asymptotically into a plane wave with exactly this direction of propagation (the arriving electron). In a thus selected solution we are further interested principally in a behavior of the "scattered" wave at infinity, for it describes the behavior of the system after the collision. We spell this out a little further. Let  $\psi_1^0(q_k), \psi_2^0(q_k), \dots$  be the eigenfunctions of the unperturbed atom (we assume that there is only a discrete spectrum). The unperturbed electron, in straight-line motion, corresponds to eigenfunctions  $\sin(2\pi/\lambda)(\alpha x + \beta y + \gamma z + \delta)$ , a continuous manifold of plane waves. Their wavelength, according to de Broglie, is connected with the energy of translation  $\tau$  by the relation  $\tau = h^2/(2\mu\lambda^2)$ . The eigenfunction of the unperturbed state in which the electron arrives from the  $+z$  direction, is thus

$$\psi_{n,\tau}^0(q_k, z) = \psi_n^0(q_k) \sin(2\pi z/\lambda).$$

Now let  $V(x, y, z; q_k)$  be the potential energy of interaction of the atom and the electron. One can then show with the help of a simple perturbation calculation that there is a uniquely determined solution of the Schrödinger equation with a potential  $V$ , which goes over asymptotically for  $z \rightarrow +\infty$  into the above function.

The question is now how this solution behaves "after the collision."

The calculation gives this result: The scattered wave created by this perturbation has asymptotically at infinity the form:

$$\psi_{nr}^1(x, y, z; q_k) = \sum_m \iint_{\alpha x + \beta y + \gamma z > 0} d\omega \Phi_{n,m}(\alpha, \beta, \gamma) \sin k_{n,m}(\alpha x + \beta y + \gamma z + \delta) \psi_m^0(q_k).$$

This means that the perturbation, analyzed at infinity, can be regarded as a superposition of solutions of the unperturbed problem. If one calculates the energy belonging to the wavelength  $\lambda_{n,m}$  according to the de Broglie formula, one finds

$$W_{n,m} = h\nu_{nm}^0 + \tau,$$

where the  $\nu_{nm}^0$  are the frequencies of the unperturbed atom.

If one translates this result into terms of particles, only one interpretation is possible.  $\Phi_{n,m}(\alpha, \beta, \gamma)$  gives the probability\* for the electron, arriving from the  $z$ -direction, to be thrown out into the direction designated by the angles  $\alpha, \beta, \gamma$ , with the phase change  $\delta$ . Here its energy  $\tau$  has increased by one quantum  $h\nu_{nm}^0$  at the cost of the energy of the atom (collision of the first kind for  $W_n^0 < W_m^0, h\nu_{nm}^0 < 0$ ; collision of the second kind  $W_n^0 > W_m^0, h\nu_{nm}^0 > 0$ ).

Schrödinger's quantum mechanics therefore gives quite a definite answer to the question of the effect of the collision; but there is no question of any causal description. One gets no answer to the question, "what is the state after the collision," but only to the question, "how probable is a specified outcome of the collision" (where naturally the quantum mechanical energy relation must be fulfilled).

Here the whole problem of determinism comes up. From the standpoint of our quantum mechanics there is no quantity which in any individual case causally fixes the consequence of the collision; but also experimentally we have so far no reason to believe that there are some inner properties of the atom which condition a definite outcome for the collision. Ought we to hope later to discover such properties (like phases or the internal atomic motions) and determine them in individual cases? Or ought we to believe that the agreement of theory and experiment—as to the impossibility of prescribing conditions for a causal evolution—is a pre-established harmony founded on the nonexistence of such conditions? I myself am inclined to give up determinism in the world of atoms. But that is a philosophical question for which physical arguments alone are not decisive.

In practical terms indeterminism is present for experimental as well as for theoretical physicists. The "yield function"  $\Phi$  so much investigated by experimentalists is now also sharply defined theoretically. One can determine it from the potential energy of interaction,  $V(x, y, z; q_k)$ . However, the calculations required

\* Addition in proof: More careful consideration shows that the probability is proportional to the square of the quantity  $\Phi_{n,m}$ .

for this purpose are too complicated to communicate here. I will only clarify briefly the meaning of the function  $\Phi_{n,m}$ . If, for example, the atom before the collision is in the normal state  $n = 1$ , then it follows from the equation

$$\tau + hv_{1m}^0 = \tau - hv_{m1}^0 = W_{1,m} > 0,$$

that, for an electron with less energy than the lowest excitation energy of the atom, the final state is also necessarily  $m = 1$ , or that  $W_{1,1}$  must be equal to  $\tau$ . Then we have "elastic reflection" of the electron with the yield function  $\Phi_{1,1}$ . If  $\tau$  increases beyond the first excitation level, then there occurs, besides reflection, also excitation with the yield  $\Phi_{1,2}$ , etc. If the target atom is in the excited state  $n = 2$  and  $\tau < hv_{21}^0$ , then there occur reflection with yield  $\Phi_{2,2}$  and collisions of the second kind with the yield  $\Phi_{2,1}$ . If the kinetic energy  $\tau > hv_{21}^0$ , then further excitation is also possible.

The formulas thus reproduce completely the qualitative character of collisions. The quantitative predictions of the formulas for particular cases require extensive investigation.

I do not exclude the possibility that the strict connection of mechanics and statistics as it comes to light here will demand a revision of basic ideas of thermodynamics and statistical mechanics.

I also believe that the problem of radiation of light—and irradiation—has to be handled in a way entirely analogous to the "boundary value problem" of the wave equation, and will lead to a rational theory of radiation damping and line-breadths in agreement with the theory of light quanta.

An extended treatment will appear shortly in this journal.

# WAVE MECHANICS

by E. Schrödinger

## Quantisation as a Problem of Proper Values (Part I)

(*Annalen der Physik* (4), vol. 79, 1926)

§ 1. In this paper I wish to consider, first, the simple case of the hydrogen atom (non-relativistic and unperturbed), and show that the customary quantum conditions can be replaced by another postulate, in which the notion of "whole numbers", merely as such, is not introduced. Rather when integralness does appear, it arises in the same natural way as it does in the case of the *node-numbers* of a vibrating string. The new conception is capable of generalisation, and strikes, I believe, very deeply at the true nature of the quantum rules.

The usual form of the latter is connected with the Hamilton-Jacobi differential equation,

$$(1) \quad H\left(q, \frac{\partial S}{\partial q}\right) = E.$$

A solution of this equation is sought such as can be represented as the sum of functions, each being a function of one only of the independent variables  $q$ .

Here we now put for  $S$  a new unknown  $\psi$  such that it will appear as a product of related functions of the single co-ordinates, i.e. we put

$$(2) \quad S = K \log \psi.$$

The constant  $K$  must be introduced from considerations of dimensions; it has those of *action*. Hence we get

$$(1') \quad H\left(q, \frac{K}{\psi} \frac{\partial \psi}{\partial q}\right) = E.$$

Now we do not look for a solution of equation (1'), but proceed as follows. If we neglect the relativistic variation of mass, equation (1') can always be transformed so as to become a quadratic form (of  $\psi$  and its first derivatives) equated to zero. (For the *one-electron* problem

this holds even when mass-variation is not neglected.) We now seek a function  $\psi$ , such that for any arbitrary variation of it the integral of the said quadratic form, taken over the whole co-ordinate space,<sup>1</sup> is stationary,  $\psi$  being everywhere real, single-valued, finite, and continuously differentiable up to the second order. *The quantum conditions are replaced by this variation problem.*

First, we will take for  $H$  the Hamilton function for Keplerian motion, and show that  $\psi$  can be so chosen for all positive, but only for a discrete set of negative values of  $E$ . That is, the above variation problem has a discrete and a continuous spectrum of proper values.

The discrete spectrum corresponds to the Balmer terms and the continuous to the energies of the hyperbolic orbits. For numerical agreement  $K$  must have the value  $h/2\pi$ .

The choice of co-ordinates in the formation of the variational equations being arbitrary, let us take rectangular Cartesians. Then (1') becomes in our case

$$(1'') \quad \left(\frac{\partial\psi}{\partial x}\right)^2 + \left(\frac{\partial\psi}{\partial y}\right)^2 + \left(\frac{\partial\psi}{\partial z}\right)^2 - \frac{2m}{K^2} \left(E + \frac{e^2}{r}\right) \psi^2 = 0;$$

$e$  = charge,  $m$  = mass of an electron,  $r^2 = x^2 + y^2 + z^2$ .

Our variation problem then reads

$$(3) \quad \delta J = \delta \iiint dx dy dz \left[ \left(\frac{\partial\psi}{\partial x}\right)^2 + \left(\frac{\partial\psi}{\partial y}\right)^2 + \left(\frac{\partial\psi}{\partial z}\right)^2 - \frac{2m}{K^2} \left(E + \frac{e^2}{r}\right) \psi^2 \right] = 0,$$

the integral being taken over all space. From this we find in the usual way

$$(4) \quad \frac{1}{2} \delta J = \int df \delta \psi \frac{\partial \psi}{\partial n} - \iiint dx dy dz \delta \psi \left[ \nabla^2 \psi + \frac{2m}{K^2} \left(E + \frac{e^2}{r}\right) \psi \right] = 0.$$

Therefore we must have, firstly,

$$(5) \quad \nabla^2 \psi + \frac{2m}{K^2} \left(E + \frac{e^2}{r}\right) \psi = 0,$$

and secondly,

$$(6) \quad \int df \delta \psi \frac{\partial \psi}{\partial n} = 0.$$

$df$  is an element of the infinite closed surface over which the integral is taken.

(It will turn out later that this last condition requires us to supplement our problem by a postulate as to the behaviour of  $\delta\psi$  at infinity, in order to ensure the existence of the above-mentioned continuous spectrum of proper values. See later.)

The solution of (5) can be effected, for example, in polar co-ordinates,  $r$ ,  $\theta$ ,  $\phi$ , if  $\psi$  be written as the product of three functions, each only of  $r$ , or of  $\theta$ , or of  $\phi$ . The method is sufficiently well known. The function of the angles turns out to be a *surface harmonic*, and if that of  $r$  be called  $\chi$ , we get easily the differential equation,

<sup>1</sup> I am aware this formulation is not entirely unambiguous.

$$(7) \quad \frac{d^2\chi}{dr^2} + \frac{2}{r} \frac{d\chi}{dr} + \left( \frac{2mE}{K^2} + \frac{2me^2}{K^2 r} - \frac{n(n+1)}{r^2} \right) \chi = 0.$$

$n = 0, 1, 2, 3 \dots$

The limitation of  $n$  to integral values is *necessary* so that the surface harmonic may be *single-valued*. We require solutions of (7) that will remain finite for all non-negative real values of  $r$ . Now<sup>1</sup> equation (7) has *two* singularities in the complex  $r$ -plane, at  $r=0$  and  $r=\infty$ , of which the second is an “indefinite point” (essential singularity) of *all* integrals, but the first on the contrary is not (for any integral). These two singularities form exactly the *bounding points of our real interval*. In such a case it is known now that the postulation of the *finiteness* of  $\chi$  at the bounding points is equivalent to a *boundary condition*. The equation has *in general* no integral which remains finite at *both* end points; such an integral exists only for certain special values of the constants in the equation. It is now a question of defining these special values. This is the *jumping-off* point of the whole investigation.<sup>2</sup>

Let us examine first the singularity at  $r=0$ . The so-called *indicial* equation which defines the behaviour of the integral at this point, is

$$(8) \quad \rho(\rho-1) + 2\rho - n(n+1) = 0,$$

with roots

$$(8') \quad \rho_1 = n, \quad \rho_2 = -(n+1).$$

The two canonical integrals at this point have therefore the exponents  $n$  and  $-(n+1)$ . Since  $n$  is not negative, only the first of these is of use to us. Since it belongs to the greater exponent, it can be represented by an ordinary power series, which begins with  $r^n$ . (The other integral, which does not interest us, can contain a logarithm, since the difference between the indices is an integer.) The next singularity is at infinity, so the above power series is always convergent and represents a *transcendental integral function*. We therefore have established that :

*The required solution is (except for a constant factor) a single-valued definite transcendental integral function, which at  $r=0$  belongs to the exponent  $n$ .*

We must now investigate the behaviour of this function at infinity on the positive real axis. To that end we simplify equation (7) by the substitution

$$(9) \quad \chi = r^\alpha U,$$

where  $\alpha$  is so chosen that the term with  $1/r^2$  drops out. It is easy to verify that then  $\alpha$  must have one of the two values  $n, -(n+1)$ . Equation (7) then takes the form,

<sup>1</sup> For guidance in the treatment of (7) I owe thanks to Hermann Weyl.

<sup>2</sup> For unproved propositions in what follows, see L. Schlesinger's *Differential Equations* (Collection Schubert, No. 13, Göschen, 1900, especially chapters 3 and 5).

$$(7') \quad \frac{d^2U}{dr^2} + \frac{2(a+1)}{r} \frac{dU}{dr} + \frac{2m}{K^2} \left( E + \frac{e^2}{r} \right) U = 0.$$

Its integrals belong at  $r=0$  to the exponents 0 and  $-2a-1$ . For the  $a$ -value,  $a=n$ , the *first* of these integrals, and for the second  $a$ -value,  $a=-(n+1)$ , the *second* of these integrals is an integral function and leads, according to (9), to the desired solution, which is single-valued. We therefore lose nothing if we confine ourselves to *one* of the two  $a$ -values. Take, then,

$$(10) \quad a=n.$$

Our solution  $U$  then, at  $r=0$ , belongs to the exponent 0. Equation (7') is called Laplace's equation. The general type is

$$(7'') \quad U'' + \left( \delta_0 + \frac{\delta_1}{r} \right) U' + \left( \epsilon_0 + \frac{\epsilon_1}{r} \right) U = 0.$$

Here the constants have the values

$$(11) \quad \delta_0 = 0, \quad \delta_1 = 2(a+1), \quad \epsilon_0 = \frac{2mE}{K^2}, \quad \epsilon_1 = \frac{2me^2}{K^2}.$$

This type of equation is comparatively simple to handle for this reason: The so-called Laplace's transformation, which in general leads *again* to an equation of the *second* order, *here* gives one of the *first*. This allows the solutions of (7'') to be represented by complex integrals. The result<sup>1</sup> only is given here. The integral

$$(12) \quad U = \int_L e^{zr} (z - c_1)^{a_1-1} (z - c_2)^{a_2-1} dz$$

is a solution of (7'') for a path of integration  $L$ , for which

$$(13) \quad \int_L \frac{d}{dz} [e^{zr} (z - c_1)^{a_1} (z - c_2)^{a_2}] dz = 0.$$

The constants  $c_1$ ,  $c_2$ ,  $a_1$ ,  $a_2$  have the following values.  $c_1$  and  $c_2$  are the roots of the quadratic equation

$$(14) \quad z^2 + \delta_0 z + \epsilon_0 = 0,$$

and

$$(14') \quad a_1 = \frac{\epsilon_1 + \delta_1 c_1}{c_1 - c_2}, \quad a_2 = -\frac{\epsilon_1 + \delta_1 c_2}{c_1 - c_2}.$$

In the case of equation (7') these become, using (11) and (10),

$$(14'') \quad c_1 = + \sqrt{\frac{-2mE}{K^2}}, \quad c_2 = - \sqrt{\frac{-2mE}{K^2}};$$

$$a_1 = \frac{me^2}{K\sqrt{-2mE}} + n + 1, \quad a_2 = -\frac{me^2}{K\sqrt{-2mE}} + n + 1.$$

+

The representation by the integral (12) allows us, not only to survey the asymptotic behaviour of the totality of solutions when  $r$

<sup>1</sup> Cf. Schlesinger. The theory is due to H. Poincaré and J. Horn.

tends to infinity in a definite way, but also to give an account of this behaviour for one *definite* solution, which is always a much more difficult task.

We shall at first *exclude* the case where  $a_1$  and  $a_2$  are real integers. When this occurs, it occurs for both quantities simultaneously, and when, and only when,

$$(15) \quad \frac{me^2}{K\sqrt{-2mE}} = \text{a real integer.}$$

+

Therefore we assume that (15) is not fulfilled.

The behaviour of the totality of solutions when  $r$  tends to infinity in a definite manner—we think always of  $r$  becoming infinite through real positive values—is characterised<sup>1</sup> by the behaviour of the two linearly independent solutions, which we will call  $U_1$  and  $U_2$ , and which are obtained by the following *specialisations* of the path of integration  $L$ . In each case let  $z$  come from infinity and return there along the same path, in such a direction that

$$(16) \quad \lim_{z \rightarrow \infty} e^{zr} = 0,$$

i.e. the real part of  $zr$  is to become negative and infinite. In this way condition (13) is satisfied. In the *one* case let  $z$  make a circuit once round the point  $c_1$  (solution  $U_1$ ), and in the *other*, round  $c_2$  (solution  $U_2$ ).

Now for very large real positive values of  $r$ , these two solutions are represented *asymptotically* (in the sense used by Poincaré) by

$$(17) \quad \begin{cases} U_1 \sim e^{c_1 r} r^{-a_1} (-1)^{a_1} (e^{2\pi i a_1} - 1) \Gamma(a_1) (c_1 - c_2)^{a_2 - 1}, \\ U_2 \sim e^{c_2 r} r^{-a_2} (-1)^{a_2} (e^{2\pi i a_2} - 1) \Gamma(a_2) (c_2 - c_1)^{a_1 - 1}, \end{cases}$$

in which we are content to take the first term of the asymptotic series of integral negative powers of  $r$ .

We have now to distinguish between the two cases.

1.  $E > 0$ . This guarantees the non-fulfilment of (15), as it makes the left hand a pure imaginary. Further, by (14''),  $c_1$  and  $c_2$  also become pure imaginaries. The exponential functions in (17), since  $r$  is real, are therefore periodic functions which remain finite. The values of  $a_1$  and  $a_2$  from (14'') show that both  $U_1$  and  $U_2$  tend to zero like  $r^{-n-1}$ . *This must therefore be valid for our transcendental integral solution  $U$ , whose behaviour we are investigating, however it may be linearly compounded from  $U_1$  and  $U_2$ .* Further, (9) and (10) show that the function  $\chi$ , i.e. the transcendental integral solution of the original equation (7), always tends to zero like  $1/r$ , as it arises from  $U$  through multiplication by  $r^n$ . We can thus state :

*The Eulerian differential equation (5) of our variation problem has, for every positive  $E$ , solutions, which are everywhere single-valued, finite, and continuous; and which tend to zero with  $1/r$  at infinity, under continual oscillations. The surface condition (6) has yet to be discussed.*

<sup>1</sup> If (15) is satisfied, at least one of the two paths of integration described in the text cannot be used, as it yields a vanishing result.

2.  $E < 0$ . In this case the possibility (15) is not *eo ipso* excluded, yet we will maintain that exclusion provisionally. Then by (14'') and (17), for  $r \rightarrow \infty$ ,  $U_1$  grows beyond all limits, but  $U_2$  vanishes exponentially. Our integral function  $U$  (and the same is true for  $\chi$ ) will then remain finite if, and only if,  $U$  is identical with  $U_2$ , save perhaps for a numerical factor. *This, however, can never be*, as is proved thus: If a closed circuit round both points  $c_1$  and  $c_2$  be chosen for the path  $L$ , thereby satisfying condition (13) since the circuit is *really closed* on the Riemann surface of the integrand, on account of  $a_1 + a_2$  being an integer, then it is easy to show that the integral (12) represents *our integral function*  $U$ . (12) can be developed in a series of positive powers of  $r$ , which converges, at all events, for  $r$  sufficiently small, and since it satisfies equation (7'), it must coincide with the series for  $U$ . Therefore  $U$  is represented by (12) if  $L$  be a closed circuit round both points  $c_1$  and  $c_2$ . This closed circuit can be so distorted, however, as to make it appear additively combined from the two paths, considered above, which belonged to  $U_1$  and  $U_2$ ; and the factors are non-vanishing, 1 and  $e^{2\pi i a_1}$ . Therefore  $U$  cannot coincide with  $U_2$ , but must contain also  $U_1$ . Q.E.D.

Our integral function  $U$ , which alone of the solutions of (7') is considered for our problem, is therefore not finite for  $r$  large, on the above hypothesis. Reserving meanwhile the question of *completeness*, *i.e.* the proving that our treatment allows us to find all the linearly independent solutions of the problem, then we may state:

*For negative values of  $E$  which do not satisfy condition (15) our variation problem has no solution.*

We have now only to investigate that discrete set of negative  $E$ -values which satisfy condition (15).  $a_1$  and  $a_2$  are then both integers. The first of the integration paths, which previously gave us the fundamental values  $U_1$  and  $U_2$ , must now undoubtedly be modified so as to give a non-vanishing result. For, since  $a_1 - 1$  is certainly positive, the point  $c_1$  is neither a branch point nor a pole of the integrand, but an ordinary zero. The point  $c_2$  can also become regular if  $a_2 - 1$  is also not negative. In *every* case, however, two suitable paths are readily found and the integration effected completely in terms of known functions, so that the behaviour of the solutions can be fully investigated.

Let

$$(15') \quad \frac{me^2}{K\sqrt{-2mE}} = l; \quad l = 1, 2, 3, 4 \dots$$

Then from (14'') we have

$$(14'') \quad a_1 - 1 = l + n, \quad a_2 - 1 = -l + n.$$

Two cases have to be distinguished:  $l \leq n$  and  $l > n$ .

(a)  $l \leq n$ . Then  $c_2$  and  $c_1$  lose every singular character, but instead become starting-points or end-points of the path of integration, in order to fulfil condition (13). A third characteristic point here is at *infinity* (negative and real). Every path between two of these three points yields a solution, and of these three solutions there are two linearly in-

dependent, as is easily confirmed if the integrals are calculated out. In particular, the *transcendental integral solution* is given by the path from  $c_1$  to  $c_2$ . That this integral remains regular at  $r=0$  can be seen at once without calculating it. I emphasize this point, as the actual calculation is apt to obscure it. However, the calculation does show that the integral becomes indefinitely great for positive, infinitely great values of  $r$ . One of the other two integrals remains *finite* for  $r$  large, but it becomes infinite for  $r=0$ .

Therefore when  $l \leq n$  we get *no* solution of the problem.

(b)  $l > n$ . Then from (14''),  $c_1$  is a zero and  $c_2$  a pole of the first order at least of the integrand. Two independent integrals are then obtained: one from the path which leads from  $z = -\infty$  to the zero, intentionally avoiding the pole; and the other from the *residue* at the pole. The latter is the integral function. We will give its calculated value, but multiplied by  $r^n$ , so that we obtain, according to (9) and (10), the solution  $\chi$  of the original equation (7). (The multiplying constant is arbitrary.) We find

$$(18) \quad \chi = f\left(r \frac{\sqrt{-2mE}}{K}\right); \quad f(x) = x^n e^{-x} \sum_{k=0}^{l-n-1} \frac{(-2x)^k}{k!} \binom{l+n}{l-n-1-k}.$$

It is seen that this is a solution that can be utilised, since it remains finite for all real non-negative values of  $r$ . In addition, it satisfies the surface condition (6) because of its vanishing exponentially at infinity. Collecting then the results for  $E$  negative:

*For  $E$  negative, our variation problem has solutions if, and only if,  $E$  satisfies condition (15). Only values smaller than  $l$  (and there is always at least one such at our disposal) can be given to the integer  $n$ , which denotes the order of the surface harmonic appearing in the equation. The part of the solution depending on  $r$  is given by (18).*

Taking into account the constants in the surface harmonic (known to be  $2n+1$  in number), it is further found that:

*The discovered solution has exactly  $2n+1$  arbitrary constants for any permissible  $(n, l)$  combination; and therefore for a prescribed value of  $l$  has  $l^2$  arbitrary constants.*

We have thus confirmed the main points of the statements originally made about the proper-value spectrum of our variation problem, but there are still deficiencies.

Firstly, we require information as to the completeness of the collected system of proper functions indicated above, but I will not concern myself with that in this paper. From experience of similar cases, it may be supposed that no proper value has escaped us.

Secondly, it must be remembered that the proper functions, ascertained for  $E$  positive, do not solve the variation problem as originally postulated, because they only tend to zero at infinity as  $1/r$ , and therefore  $\delta\psi/\delta r$  only tends to zero on an infinite sphere as  $1/r^2$ . Hence the surface integral (6) is still of the same order as  $\delta\psi$  at infinity. If it is desired therefore to obtain the continuous spectrum, another condition must be added to the *problem*, viz. that  $\delta\psi$  is to vanish at

infinity, or at least, that it tends to a constant value independent of the direction of proceeding to infinity; in the latter case the surface harmonics cause the surface integral to vanish.

§ 2. Condition (15) yields

$$(19) \quad -E_l = \frac{me^4}{2K^2l^2}.$$

Therefore the well-known Bohr energy-levels, corresponding to the Balmer terms, are obtained, if to the constant  $K$ , introduced into (2) for reasons of dimensions, we give the value

$$(20) \quad K = \frac{\hbar}{2\pi},$$

from which comes

$$(19') \quad -E_l = \frac{2\pi^2 me^4}{\hbar^2 l^2}.$$

Our  $l$  is the principal quantum number.  $n+1$  is analogous to the azimuthal quantum number. The splitting up of this number through a closer definition of the surface harmonic can be compared with the resolution of the azimuthal quantum into an "equatorial" and a "polar" quantum. These numbers *here* define the system of node-lines on the sphere. Also the "radial quantum number"  $l-n-1$  gives exactly the number of the "node-spheres", for it is easily established that the function  $f(x)$  in (18) has exactly  $l-n-1$  positive real roots. The positive  $E$ -values correspond to the continuum of the hyperbolic orbits, to which one may ascribe, in a certain sense, the radial quantum number  $\infty$ . The fact corresponding to this is the proceeding to infinity, under *continual* oscillations, of the functions in question.

It is interesting to note that the range, inside which the functions of (18) differ sensibly from zero, and outside which their oscillations die away, is of the *general order of magnitude* of the major axis of the ellipse in each case. The factor, multiplied by which the radius vector enters as the argument of the constant-free function  $f$ , is—naturally—the reciprocal of a length, and this length is

$$(21) \quad \frac{K}{\sqrt{-2mE}} = \frac{K^2 l}{me^2} = \frac{\hbar^2 l}{4\pi^2 me^2} = \frac{a_l}{l},$$

where  $a_l$  = the semi-axis of the  $l$ th elliptic orbit. (The equations follow from (19) plus the known relation  $E_l = \frac{-e^2}{2a_l}$ ).

The quantity (21) gives the order of magnitude of the range of the roots when  $l$  and  $n$  are small; for then it may be assumed that the roots of  $f(x)$  are of the order of unity. That is naturally no longer the case if the coefficients of the polynomial are large numbers. At present I will not enter into a more exact evaluation of the roots, though I believe it would confirm the above assertion pretty thoroughly.

§ 3. It is, of course, strongly suggested that we should try to connect the function  $\psi$  with some *vibration process* in the atom, which would more nearly approach reality than the electronic orbits, the real existence of which is being very much questioned to-day. I originally intended to find the new quantum conditions in this more intuitive manner, but finally gave them the above neutral mathematical form, because it brings more clearly to light what is really essential. The essential thing seems to me to be, that the postulation of "whole numbers" no longer enters into the quantum rules mysteriously, but that we have traced the matter a step further back, and found the "integralness" to have its origin in the finiteness and single-valuedness of a certain space function.

I do not wish to discuss further the possible representations of the vibration process, before more complicated cases have been calculated successfully from the new stand-point. It is not decided that the results will merely re-echo those of the usual quantum theory. For example, if the relativistic Kepler problem be worked out, it is found to lead in a remarkable manner to *half-integral partial quanta* (radial and azimuthal).

Still, a few remarks on the representation of the vibration may be permitted. Above all, I wish to mention that I was led to these deliberations in the first place by the suggestive papers of M. Louis de Broglie,<sup>1</sup> and by reflecting over the space distribution of those "phase waves", of which he has shown that there is always a *whole number*, measured along the path, present on each period or quasi-period of the electron. The main difference is that de Broglie thinks of progressive waves, while we are led to stationary proper vibrations if we interpret our formulae as representing vibrations. I have lately shown<sup>2</sup> that the Einstein gas theory can be based on the consideration of such stationary proper vibrations, to which the dispersion law of de Broglie's phase waves has been applied. The above reflections on the atom could have been represented as a generalisation from those on the gas model.

If we take the separate functions (18), multiplied by a surface harmonic of order  $n$ , as the description of proper vibration processes, then the quantity  $E$  must have something to do with the related *frequency*. Now in vibration problems we are accustomed to the "parameter" (usually called  $\lambda$ ) being proportional to the *square* of the frequency. However, in the first place, such a statement in our case would lead to *imaginary* frequencies for the *negative E-values*, and, secondly, instinct leads us to believe that the energy must be proportional to the frequency itself and not to its square.

The contradiction is explained thus. There has been *no natural zero level* laid down for the "parameter"  $E$  of the variation equation (5), especially as the unknown function  $\psi$  appears multiplied by a function of  $r$ , which can be changed by a constant to meet a corresponding

<sup>1</sup> L. de Broglie, *Ann. de Physique* (10) 3, p. 22, 1925. (*Thèses, Paris, 1924.*)

<sup>2</sup> *Physik. Ztschr.* 27, p. 95, 1926.

change in the zero level of  $E$ . Consequently, we have to correct our anticipations, in that not  $E$  itself—continuing to use the same terminology—but  $E$  increased by a certain constant is to be expected to be proportional to the square of the frequency. Let this constant be now *very great* compared with all the admissible negative  $E$ -values (which are already limited by (15)). Then firstly, the frequencies will become *real*, and secondly, since our  $E$ -values correspond to only relatively small frequency *differences*, they will actually be very approximately proportional to these frequency differences. This, again, is all that our “quantum-instinct” can require, as long as the zero level of energy is not fixed.

The view that the frequency of the vibration process is given by

$$(22) \quad \nu = C' \sqrt{C + E} = C' \sqrt{C} + \frac{C'}{2\sqrt{C}} E + \dots,$$

where  $C$  is a constant very great compared with all the  $E$ 's, has still another very appreciable advantage. *It permits an understanding of the Bohr frequency condition.* According to the latter the emission frequencies are proportional to the  $E$ -differences, and therefore from (22) also to the differences of the proper frequencies  $\nu$  of those hypothetical vibration processes. But these proper frequencies are all very great compared with the emission frequencies, and they agree very closely among themselves. The emission frequencies appear therefore as deep “difference tones” of the proper vibrations themselves. It is quite conceivable that on the transition of energy from one to another of the normal vibrations, *something*—I mean the light wave—with a *frequency* allied to each frequency *difference*, should make its appearance. One only needs to imagine that the light wave is causally related to the *beats*, which necessarily arise at each point of space during the transition; and that the frequency of the light is defined by the number of times per second the intensity maximum of the beat-process repeats itself.

It may be objected that these conclusions are based on the relation (22), in its *approximate* form (after expansion of the square root), from which the Bohr frequency condition itself seems to obtain the nature of an approximation. This, however, is merely apparently so, and it is wholly avoided when the *relativistic* theory is developed and makes a profounder insight possible. The large constant  $C$  is naturally very intimately connected with the rest-energy of the electron ( $mc^2$ ). Also the seemingly *new* and *independent* introduction of the constant  $h$  (already brought in by (20)), into the frequency condition, is cleared up, or rather avoided, by the relativistic theory. But unfortunately the correct establishment of the latter meets right away with certain difficulties, which have been already alluded to.

- It is hardly necessary to emphasize how much more congenial it would be to imagine that at a quantum transition the energy changes over from one form of vibration to another, than to think

of a jumping electron. The changing of the vibration form can take place continuously in space and time, and it can readily last as long as the emission process lasts empirically (experiments on canal rays by W. Wien); nevertheless, if during this transition the atom is placed for a comparatively short time in an electric field which alters the proper frequencies, then the beat frequencies are immediately changed sympathetically, and for just as long as the field operates. It is known that this experimentally established fact has hitherto presented the greatest difficulties. See the well-known attempt at a solution by Bohr, Kramers, and Slater.

Let us not forget, however, in our gratification over our progress in these matters, that the idea of only *one* proper vibration being excited whenever the atom does not radiate—if we must hold fast to this idea—is very far removed from the *natural* picture of a vibrating system. We know that a macroscopic system does not behave like that, but yields in general a *pot-pourri* of its proper vibrations. But we should not make up our minds too quickly on this point. A *pot-pourri* of proper vibrations would also be permissible for a single atom, since thereby no beat frequencies could arise other than those which, according to experience, the atom is capable of emitting *occasionally*. The actual sending out of many of these spectral lines simultaneously by the same atom does not contradict experience. It is thus conceivable that only in the normal state (and approximately in certain “meta-stable” states) the atom vibrates with *one* proper frequency and just for this reason does *not* radiate, namely, because no beats arise. The *stimulation* may consist of a simultaneous excitation of one or of several other proper frequencies, whereby beats originate and evoke emission of light.

Under all circumstances, I believe, the proper functions, which belong to the *same* frequency, are in general all simultaneously stimulated. Multiplicity of the proper values corresponds, namely, in the language of the previous theory to *degeneration*. To the reduction of the quantisation of degenerate systems probably corresponds the arbitrary partition of the energy among the functions belonging to *one* proper value.

*Addition at the proof correction on 28.2.1926.*

In the case of conservative systems in classical mechanics, the variation problem can be formulated in a neater way than was previously shown, and without express reference to the Hamilton-Jacobi differential equation. Thus, let  $T(q, p)$  be the kinetic energy, expressed as a function of the co-ordinates and momenta,  $V$  the potential energy, and  $d\tau$  the volume element of the space, “measured rationally”, i.e. it is not simply the product  $dq_1 dq_2 dq_3 \dots dq_n$ , but this divided by the square root of the discriminant of the quadratic form  $T(q, p)$ . (Cf. Gibbs’ *Statistical Mechanics*.) Then let  $\psi$  be such as to make the “Hamilton integral”

$$(23) \quad \int d\tau \left\{ K^2 T \left( q, \frac{\partial \psi}{\partial q} \right) + \psi^2 V \right\}$$

*stationary, while fulfilling the normalising, accessory condition*

$$(24) \quad \int \psi^2 d\tau = 1.$$

The proper values of this variation problem are then *the stationary values of integral (23)* and yield, according to our thesis, *the quantum-levels of the energy.*

It is to be remarked that in the quantity  $a_2$  of (14'') we have essentially the well-known Sommerfeld expression  $-\frac{B}{\sqrt{A}} + \sqrt{C}$ . (Cf. *Atombau*, 4th (German) ed., p. 775.)

Physical Institute of the University of Zürich.

(Received January 27, 1926.)

### I.3 THE PHYSICAL CONTENT OF QUANTUM KINEMATICS AND MECHANICS

WERNER HEISENBERG

First we define the terms *velocity*, *energy*, etc. (for example, for an electron) which remain valid in quantum mechanics. It is shown that canonically conjugate quantities can be determined simultaneously only with a characteristic indeterminacy (§1). This indeterminacy is the real basis for the occurrence of statistical relations in quantum mechanics. Its mathematical formulation is given by the Dirac-Jordan theory (§2). Starting from the basic principles thus obtained, we show how microscopic processes can be understood by way of quantum mechanics (§3). To illustrate the theory, a few special *gedankenexperiments* are discussed (§4).

We believe we understand the physical content of a theory when we can see its qualitative experimental consequences in all simple cases and when at the same time we have checked that the application of the theory never contains inner contradictions. For example, we believe that we understand the physical content of Einstein's concept of a closed 3-dimensional space because we can visualize consistently the experimental consequences of this concept. Of course these consequences contradict our everyday physical concepts of space and time. However, we can convince ourselves that the possibility of employing usual space-time concepts at cosmological distances can be justified neither by logic nor by observation. The physical interpretation of quantum mechanics is still full of internal discrepancies, which show themselves in arguments about continuity versus discontinuity and particle versus wave. Already from this circumstance one might conclude that no interpretation of quantum mechanics is possible which uses ordinary kinematical and mechanical concepts. Of course, quantum mechanics arose exactly out of the attempt to break with all ordinary kinematic concepts and to put in their place relations between concrete and experimentally determinable numbers. Moreover, as this enterprise seems to have succeeded, the mathematical scheme of quantum mechanics needs no revision. Equally unnecessary is a revision of space-time geometry at small distances, as we can make the quantum-mechanical laws approximate the classical ones arbitrarily closely by choosing sufficiently great masses, even when arbitrarily small distances and times come into question. But that a revision of kinematical and mechanical concepts is necessary

seems to follow directly from the basic equations of quantum mechanics. When a definite mass  $m$  is given, in our everyday physics it is perfectly understandable to speak of the position and the velocity of the center of gravity of this mass. In quantum mechanics, however, the relation  $\mathbf{pq} - \mathbf{qp} = -i\hbar$  between mass, position, and velocity is believed to hold. Therefore we have good reason to become suspicious every time uncritical use is made of the words "position" and "velocity." When one admits that discontinuities are somehow typical of processes that take place in small regions and in short times, then a contradiction between the concepts of "position" and "velocity" is quite plausible. If one considers, for example, the motion of a particle in one dimension, then in continuum theory one will be able to draw (Fig. 1) a worldline  $x(t)$  for the track of the particle (more precisely, its center of gravity), the tangent of which gives the velocity at every instant. In contrast, in a theory based on discontinuity there might be in place of this curve a series of points at finite separation (Fig. 2). In this case it is clearly meaningless to speak about one velocity at one position (1) because one velocity can only be defined by two positions and (2), conversely, because any one point is associated with two velocities.

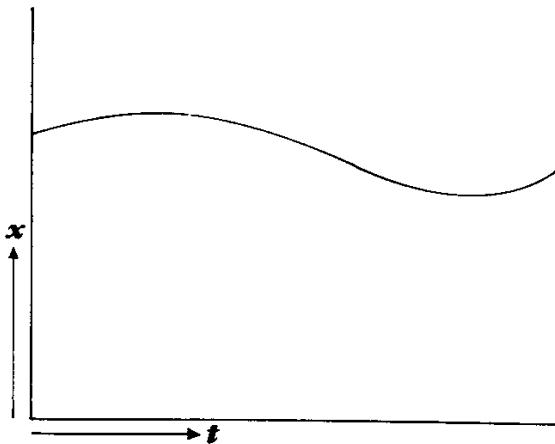


FIGURE 1

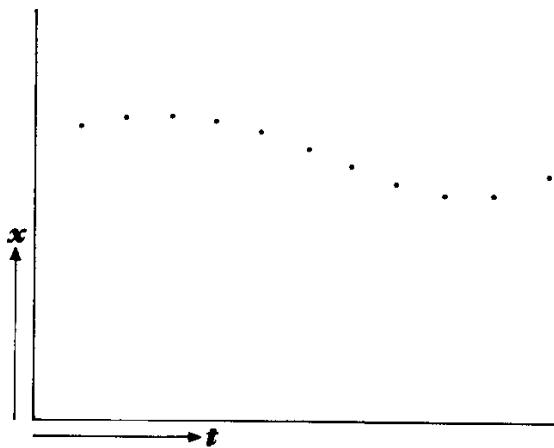


FIGURE 2

The question therefore arises whether, through a more precise analysis of these kinematic and mechanical concepts, it might be possible to clear up the contradictions evident up to now in the physical interpretations of quantum mechanics and to arrive at a physical understanding of the quantum-mechanical formulas.\*

\* The present work has arisen from efforts and desires to which other investigators have already given clear expression, before the development of quantum mechanics. I call attention here especially to Bohr's papers on the basic postulates of quantum theory (for example, *Zeits. f. Physik*, 13, 117 [1923]) and Einstein's discussions on the connection between wave field and light quanta. The problems dealt with here are discussed most clearly in recent times, and the problems arising are partly answered, by W. Pauli ("Quantentheorie," *Handbuch der Physik*, Vol. XXIII, cited hereafter as *I.c.*); quantum mechanics has changed only slightly the formulation of these problems as given by Pauli. It is also a special pleasure to thank here Herrn Pauli for the repeated stimulus I have received from our oral and written discussions, which have contributed decisively to the present work.  
814

In order to be able to follow the quantum-mechanical behavior of any object one has to know the mass of this object and its interactions with any fields and other objects. Only then can the Hamiltonian function be written down for the quantum-mechanical system. (The following considerations ordinarily refer to nonrelativistic quantum mechanics, as the laws of quantum electrodynamics are still very incompletely known.)\* About the "Gestalt" (construction) of the object any further assumption is unnecessary; one most usefully employs the word "Gestalt" to designate the totality of these interactions.

When one wants to be clear about what is to be understood by the words "position of the object," for example of the electron (relative to a given frame of reference), then one must specify definite experiments with whose help one plans to measure the "position of the electron"; otherwise this word has no meaning. There is no shortage of such experiments, which in principle even allow one to determine the "position of the electron" with arbitrary accuracy. For example, let one illuminate the electron and observe it under a microscope. Then the highest attainable accuracy in the measurement of position is governed by the wavelength of the light. However, in principle one can build, say, a  $\gamma$ -ray microscope and with it carry out the determination of position with as much accuracy as one wants. In this measurement there is an important feature, the Compton effect. Every observation of scattered light coming from the electron presupposes a photoelectric effect (in the eye, on the photographic plate, in the photocell) and can therefore also be so interpreted that a light quantum hits the electron, is reflected or scattered, and then, once again bent by the lens of the microscope, produces the photoeffect. At the instant when position is determined—therefore, at the moment when the photon is scattered by the electron—the electron undergoes a discontinuous change in momentum. This change is the greater the smaller the wavelength of the light employed—that is, the more exact the determination of the position. At the instant at which the position of the electron is known, its momentum therefore can be known up to magnitudes which correspond to that discontinuous change. Thus, the more precisely the position is determined, the less precisely the momentum is known, and conversely. In this circumstance we see a direct physical interpretation of the equation  $\mathbf{pq} - \mathbf{qp} = -i\hbar$ . Let  $q_1$  be the precision with which the value  $q$  is known ( $q_1$  is, say, the mean error of  $q$ ), therefore here the wavelength of the light. Let  $p_1$  be the precision with which the value  $p$  is determinable; that is, here, the discontinuous change of  $p$  in the Compton effect. Then, according to the elementary laws of the Compton effect  $p_1$  and  $q_1$  stand in the relation

\* Quite recently, however, great advances in this domain have been made in the papers of P. Dirac [*Proc. Roy. Soc. A114*, 243 (1927) and papers to appear subsequently].

$$p_1 q_1 \sim h. \quad (1)$$

That this relation (1) is a straightforward mathematical consequence of the rule  $\mathbf{pq} - \mathbf{qp} = -i\hbar$  will be shown below. Here we can note that equation (1) is a precise expression for the facts which one earlier sought to describe by the division of phase space into cells of magnitude  $h$ . For the determination of the position of the electron one can also do other experiments—for example, collision experiments. A precise measurement of the position demands collisions with very fast particles, because for slow electrons the diffraction phenomena—which, according to Einstein, are consequences of de Broglie waves (as, for example, in the Ramsauer effect)—prevent a sharp specification of location. In a precise measurement of position the momentum of the electron again changes discontinuously. An elementary estimate of the precision using the formulas for de Broglie waves leads once more to relation (1).

Throughout this discussion the concept of “position of the electron” seems well enough defined, and only a word need be added about the “size” of the electron. When two very fast particles hit the electron one after the other within a very short time interval  $\Delta t$ , then the positions of the electron defined by the two particles lie very close together at a distance  $\Delta l$ . From the regularities which are observed for  $\alpha$ -particles we conclude that  $\Delta l$  can be pushed down to a magnitude of the order of  $10^{-12}$  cm if only  $\Delta t$  is sufficiently small and particles are selected with sufficiently great velocity. This is what we mean when we say that the electron is a corpuscle whose radius is not greater than  $10^{-12}$  cm.

We turn now to the concept of “path of the electron.” By path we understand a series of points in space (in a given reference system) which the electron takes as “positions” one after the other. As we already know what is to be understood by “position at a definite time,” no new difficulties occur here. Nevertheless, it is easy to recognize that, for example, the often used expression, the “1s orbit of the electron in the hydrogen atom,” from our point of view has no sense. In order to measure this 1s “path” we have to illuminate the atom with light whose wavelength is considerably shorter than  $10^{-8}$  cm. However, a single photon of such light is enough to eject the electron completely from its “path” (so that only a single point of such a path can be defined). Therefore here the word “path” has no definable meaning. This conclusion can already be deduced, without knowledge of the recent theories, simply from the experimental possibilities.

In contrast, the contemplated measurement of position can be carried out on many atoms in a 1s state. (In principle, atoms in a given “stationary” state can be selected, for example, by the Stern-Gerlach experiment.) There must therefore exist for a definite state—for example, the 1s state—of the atom a probability function for the location of the electron which corresponds to the mean value for 81the classical orbit, averaged over all phases, and which can be determined through

the measurement with an arbitrary precision. According to Born,\* this function is given by  $\psi_{1s}(q)\bar{\psi}_{1s}(q)$  where  $\psi_{1s}(q)$  designates the Schrödinger wave function belonging to the 1s state. With a view to later generalizations I should like to say—with Dirac and Jordan—that the probability is given by  $S(1s, q)\bar{S}(1s, q)$ , where  $S(1s, q)$  designates that column of the matrix  $S(E, q)$  of transformation from  $E$  to  $q$  that belongs to the energy  $E = E_{1s}$ .

In the fact that in quantum theory only the probability distribution of the position of the electrons can be given for a definite state, such as 1s, one can recognize, with Born and Jordan, a characteristically statistical feature of quantum theory as contrasted to classical theory. However, one can say, if one will, with Dirac, that the statistics are brought in by our experiments. For plainly *even in classical theory* only the probability of a definite position for the electron can be given as long as we do not know the phase of [the motion of the electron in] the atom. The distinction between classical and quantum mechanics consists rather in this: classically we can always think of the phase as determined through suitable experiments. In reality, however, this is impossible, because every experiment for the determination of phase perturbs or changes the atom. In a definite stationary “state” of the atom, the phases are in principle indeterminate, as one can see as a direct consequence of the familiar equations

$$Et - tE = -i\hbar \quad \text{or} \quad Jw - wJ = -i\hbar,$$

where  $J$  is the action variable and  $w$  is the angle variable.

The word “velocity” can easily be defined for an object by measurements when the motion is free of force. For example, one can illuminate the object with red light and by way of the Doppler effect in the scattered light determine the velocity of the particle. The determination of the velocity is the more exact the longer the wavelength of the light that is used, as then the change in velocity of the particle, per light quantum, by way of the Compton effect is so much less. The determination of position becomes correspondingly inexact, in agreement with equation (1). If one wants to measure the velocity of the electron in the atom at a definite instant, then, for example, one will let the nuclear charge and the forces arising

\* The statistical interpretation of de Broglie waves was first formulated by A. Einstein (*Sitzungsber. d. preussische Akad. d. Wiss.*, p. 3 [1925]). This statistical feature of quantum mechanics then played an essential role in M. Born, W. Heisenberg, and P. Jordan, Quantenmechanik II (*Zeits. f. Physik*, 35, 557 [1926]), especially chapter 4, §3, and P. Jordan (*Zeits. f. Physik*, 37, 376 [1926]). It was analyzed mathematically in a seminal paper of M. Born (*Zeits. f. Physik*, 38, 803 [1926]) and used for the interpretation of collision phenomena. One finds how to base the probability picture on the theory of the transformation of matrices in the following papers: W. Heisenberg (*Zeits. f. Physik*, 40, 501 [1926]), P. Jordan (*Zeits. f. Physik*, 40, 661 [1926]), W. Pauli (remark in *Zeits. f. Physik*, 41, 81 [1927]), P. Dirac (*Proc. Roy. Soc. A113*, 621 [1926]), and P. Jordan (*Zeits. f. Physik*, 40, 809 [1926]). The statistical side of quantum mechanics is discussed more generally in P. Jordan (*Naturwiss.*, 15, 105 [1927]) and M. Born (*Naturwiss.*, 15, 238 [1927]).

from the other electrons suddenly be taken away, so that the motion from then on is force-free, and one will then carry out the measurement described above. As above, one can again easily convince oneself that a [momentum] function  $p(t)$  cannot be defined for a given state—such as the 1s-state—of an atom. On the contrary, there is again a probability function for  $p$  in this state which according to Dirac and Jordan has the value  $S(1s, p)\bar{S}(1s, p)$ . Here  $S(1s, p)$  again designates that column of the matrix  $S(E, p)$ —that transforms from  $\mathbf{E}$  to  $\mathbf{p}$ —which belongs to  $E = E_{1s}$ .

Finally we come to experiments which allow one to measure the energy or the value of the action variable  $J$ . Such experiments are especially important because only with their help can we define what we mean when we speak of the discontinuous change of the energy and of  $J$ . The Franck-Hertz collision experiments allow one to base the measurement of the energy of the atom on the measurement of the energy of electrons in rectilinear motion, because of the validity of the law of conservation of energy in quantum theory. This measurement in principle can be carried out with arbitrary accuracy if only one forgoes the simultaneous determination of the position of the electron or its phase (see the determination of  $p$ , above), corresponding to the relation  $\mathbf{Et} - \mathbf{tE} = -i\hbar$ . The Stern-Gerlach experiment allows one to determine the magnetic or an average electric moment of the atom, and therefore to measure quantities which depend only on the action variable  $J$ . The phases remain undetermined in principle. It makes as little sense to speak of the frequency of the light wave at a definite instant as of the energy of an atom at a definite moment. Correspondingly, in the Stern-Gerlach experiment the accuracy of the energy measurement decreases as we shorten the time during which the atom is under the influence of the deflecting field.\* Specifically, an upper bound is given for the deviating force through the circumstance that the potential energy of that deflecting force can at most vary inside the beam by an amount which is considerably smaller than the differences in energy of the stationary states. Only then will a determination of the energy of the stationary states be at all possible. Let  $E_1$  be an amount of energy which satisfies this condition ( $E_1$  also fixes the precision of the energy measurement). Then  $E_1/d$  specifies the highest allowable value for the deflecting force, if  $d$  is the breadth of the beam (measurable through the spacing of the slits employed). The angular deviation of the atomic beam is then  $E_1 t_1 / dp$ , where we designate by  $t_1$  the time during which the atoms are under the influence of the deflecting field, and by  $p$  the momentum of the atoms in the direction of the beam. This deflection must be of at least the same order of magnitude as the natural broadening of the beam brought about by the diffraction by the slits, if any measurement is to be possible. The diffraction angle is roughly  $\lambda/d$  if  $\lambda$  denotes the de Broglie wavelength; thus,

818 \* In this connection see W. Pauli, *I.c.*, p. 61.

$$\lambda/d \sim E_1 t_1/dp,$$

or, as  $\lambda = h/p$ ,

$$E_1 t_1 \sim h. \quad (2)$$

This equation corresponds to equation (1) and shows how a precise determination of energy can only be obtained at the cost of a corresponding uncertainty in the time.

## §2. THE DIRAC-JORDAN THEORY

We might summarize and generalize the results of the preceding section in this statement: *All concepts which can be used in classical theory for the description of a mechanical system can also be defined exactly for atomic processes in analogy to the classical concepts.* The experiments which provide such a definition themselves suffer an indeterminacy introduced purely by the observational procedures we use when we ask of them the simultaneous determination of two canonically conjugate quantities. The magnitude of this indeterminacy is given by relation (1) (generalized to any canonically conjugate quantities whatsoever). It is natural in this respect to compare quantum theory with special relativity. According to relativity, the word "simultaneous" cannot be defined except through experiments in which the velocity of light enters in an essential way. If there existed a "sharper" definition of simultaneity, as, for example, signals which propagate infinitely fast, then relativity theory would be impossible. However, because there are no such signals, or, rather, because already in the definition of simultaneity the velocity of light appears, there is room left for the postulate of the constancy of the speed of light so that this postulate does not contradict any meaningful use of the words "position, velocity, time." We find a similar situation with the definition of the concepts of "position of an electron" and "velocity" in quantum theory. All experiments which we can use for the definition of these terms necessarily contain the uncertainty implied by equation (1), even though they permit one to define exactly the concepts  $p$  and  $q$  taken in isolation. If there existed experiments which allowed simultaneously a "sharper" determination of  $p$  and  $q$  than equation (1) permits, then quantum mechanics would be impossible. Thus only the uncertainty which is specified by equation (1) creates room for the validity of the relations which find their most pregnant expression in the quantum-mechanical commutation relations,

$$\mathbf{pq} - \mathbf{qp} = -i\hbar.$$

That uncertainty makes possible this equation without requiring that the physical meaning of the quantities  $p$  and  $q$  be changed. 819

For those physical phenomena whose quantum-mechanical formulation is still

unknown (for example, electrodynamics), equation (1) makes a demand which may be useful for the discovery of the new laws. For quantum mechanics equation (1) can be derived from the Dirac-Jordan formulation by a slight generalization. If, for any definite state variable  $\eta$  we determine the position  $q$  of the electron as  $q'$  with an uncertainty  $q_1$ , then we can express this fact by a probability amplitude  $S(\eta, q)$  which differs appreciably from zero only in a region of spread  $q_1$  near  $q'$ . For example, one can write

$$S(\eta, q) \text{ proportional to } \exp[-(q - q')^2/2q_1^2 - (ip'/\hbar)(q - q')], \quad (3a)$$

with therefore

$$S\bar{S}' \text{ proportional to } \exp[-(q - q')^2/q_1^2]. \quad (3b)$$

Then for the probability amplitude for any given value of  $p$  we have

$$S(\eta, p) = \int S(\eta, q)S(q, p) dq. \quad (4)$$

For  $S(q, p)$ , according to Jordan, we can write

$$S(q, p) = \exp(ipq/\hbar). \quad (5)$$

Then, according to (4),  $S(\eta, p)$  differs appreciably from zero only for values of  $p$  for which  $(p - p')q_1/\hbar$  is not significantly greater than 1. Specifically, employing (3), we find  $S(\eta, p)$  is proportional to

$$\int \exp[i(p - p')q/\hbar - (q - q')^2/2q_1^2] dq;$$

that is, proportional to

$$\exp[-(p - p')^2/2p_1^2 + iq'(p - p')/\hbar];$$

and thus

$$S\bar{S} \text{ is proportional to } \exp[-(p - p')^2/p_1^2],$$

where

$$p_1 q_1 = \hbar. \quad (6)$$

The assumption (3) for  $S(\eta, q)$  corresponds therefore to the experimental fact that the value  $p'$  is measured for  $p$  and the value  $q'$  for  $q$  [with the limit (6) on the precision].

From the purely mathematical point of view it is characteristic of the Dirac-

Jordan formulation of quantum mechanics that the relations between  $\mathbf{p}$ ,  $\mathbf{q}$ ,  $\mathbf{E}$ , etc. can be described as equations connecting very general matrices in such a way that any predetermined quantum-theoretic quantity appears as a diagonal matrix. The possibility of writing things in such a way is evident when one pictures the matrices as tensors (for example, moment-of-inertia tensors) in a multidimensional space between which there are mathematical connections. One can always pick the axis of the coordinate system in which one expresses these relations along the principal axes of one of these tensors. Finally, one also can always characterize the mathematical relation between two tensors  $A$  and  $B$  through the transformation equations which take a coordinate system oriented along the principal axes of  $A$  over into another oriented along the principal axes of  $B$ . This latter formulation corresponds to the Schrödinger theory. In contrast, one will view Dirac's  $q$ -number formulation as the formulation of quantum mechanics that is really "invariant" and independent of all coordinate systems. When we want to derive physical results from that mathematical framework, then we have to associate numbers with the quantum-theoretical magnitudes—that is, with the matrices (or "tensors" in multidimensional space). This task is to be understood in these terms: In that multidimensional space a definite direction is arbitrarily prescribed (by the nature of the experimental setup); and it is asked what is the "value" of the matrix (for example, in that picture, what is the value of the moment of inertia) in this given direction. This question only has a well-defined meaning when the given direction coincides with the direction of one of the principal axes of that matrix. In this case there is an exact answer for the question. But also when the prescribed direction differs only little from one of the principal axes of the matrix one can still speak of the "value" of the matrix in the prescribed direction up to a definite uncertainty determined by the angle between the two directions. One can therefore say that associated with every quantum-theoretical quantity or matrix is a number which gives its "value" within a certain definite statistical error. The statistical error depends on the coordinate system. For every quantum-theoretical quantity there exists a coordinate system in which the statistical error for this quantity is zero. Therefore a definite experiment can never give exact information on all quantum-theoretical quantities. Rather, it divides physical quantities into "known" and "unknown" (or more and less accurately known quantities) in a way characteristic of the experiment in question. The results of two experiments can be derived exactly one from the other only then when the two experiments divide the physical quantities in the same way into "known" and "unknown" (that is, when the tensors in that multidimensional space frequently invoked—for ease of visualization—are "looked at" in both experiments from the same direction). When two experiments use different divisions into "known" and "unknown," then their results can be related only statistically.

For a more detailed discussion of this statistical connection let a *gedanke*<sup>821</sup> *experiment* be considered. Let a Stern-Gerlach atomic beam be sent first through

a field  $F_1$  which is so strongly inhomogeneous in the direction of the beam that it induces many transitions by sudden reversal in the force on the spin. Then let the atomic beam run free up to a definite distance from  $F_1$ . But there let a second field  $F_2$  begin, as inhomogeneous as  $F_1$ . Between  $F_1$  and  $F_2$  let it be possible to measure the number of atoms in the different stationary states through an optionally applied magnetic field. Let all radiation by the atoms be neglected. If we know that an atom was in a state of energy  $E_n$  before it passed  $F_1$ , then we can express this fact by ascribing to the atom a wave function—for example, in  $p$ -space—with the definite energy  $E_n$  and the undetermined phase  $\beta_n$ ,

$$S(E_n, p) = \psi(E_n, p) \exp[-i(E_n/\hbar)(\alpha + \beta_n)].$$

After passage through the field  $F_1$ , this function is changed into\*

$$S(E_n, p) \xrightarrow{F_1} \sum_m c_{nm} \psi(E_m, p) \exp[-i(E_m/\hbar)(\alpha + \beta_m)]. \quad (7)$$

Here we can make some arbitrary determination of the  $\beta_m$  so that the  $c_{nm}$  are uniquely determined by  $F_1$ . The matrix  $c_{nm}$  transforms the energy values before the transition through  $F_1$  to the values after the transition. If after  $F_1$  we carry out a determination of the stationary state, say, by use of an inhomogeneous magnetic field, then we will find that the atom has jumped from the  $n$ th state to the  $m$ th state with a probability  $c_{nm}\bar{c}_{nm}$ . When we find experimentally that an atom has indeed jumped to the  $m$ th state, then we have to ascribe to it in all calculations thereafter, not the function  $\sum_m c_{nm} S_m$ , but simply the function  $S_m$  with an undetermined phase. Through the experimental determination, “ $m$ th state,” we select out of the multitude of different possibilities ( $c_{nm}$ ) a definite one,  $m$ . However, at the same time we disturb everything that was still contained in the phase relations between the quantities  $c_{nm}$ , as detailed below. In the transition of the atomic beam through  $F_2$ , what happened at  $F_1$  repeats itself. Let  $d_{ml}$  be the coefficients of the transformation matrix which transform the energies before  $F_2$  to the energies after  $F_2$ . If no determination of the state is carried out between  $F_1$  and  $F_2$ , then the eigenfunction is transformed according to the following scheme,

$$S(E_n, p) \xrightarrow{F_1} \sum_m c_{nm} S(E_m, p) \xrightarrow{F_2} \sum_m \sum_l c_{nm} d_{ml} S(E_l, p). \quad (8)$$

Let  $\sum_m c_{nm} d_{ml}$  be called  $e_{nl}$ . If the stationary state of the atom is determined beyond  $F_2$ , then one will find the state  $l$  with the probability  $e_{nl}\bar{e}_{nl}$ . In contrast, if between  $F_1$  and  $F_2$  one determines the state—and finds for it the value  $E_m$ —then

\* See P. Dirac (*Proc. Roy. Soc. A112*, 661 [1926]) and M. Born (*Zeits. f. Physik*, **40**, 167 [1926]).

the probability for " $l$ " beyond  $F_2$  is given by  $d_{ml}\bar{d}_{ml}$ . In many repetitions of the entire experiment (in which each time the state is determined between  $F_1$  and  $F_2$ ) one will therefore observe the state  $l$ , beyond  $F_2$ , with the relative frequency  $Z_{nl} = \sum_m c_{nm}\bar{c}_{nm}d_{ml}\bar{d}_{ml}$ . This expression does not agree at all with  $e_{nl}\bar{e}_{nl}$ . For this reason

Jordan (*l.c.*) has spoken of an "interference of probabilities." I cannot agree. The two kinds of experiments which lead respectively to  $e_{nl}\bar{e}_{nl}$  and  $Z_{nl}$  are physically distinct. In one case the atom experiences no disturbance between  $F_1$  and  $F_2$ . In the other case it is perturbed by the apparatus which determines its stationary state. This apparatus has as a consequence that the "phase" of the atom changes by an amount that is in principle uncheckable, as the momentum of an electron likewise changes with a determination of its position (see §1). The magnetic field for the determination of the state between  $F_1$  and  $F_2$  will separate the eigenvalues  $E$ . In the observation of the path of the atomic beam the atoms are slowed down by statistically different and uncheckable amounts (I think here, say, of Wilson cloud-chamber pictures). This has as a consequence that the final transformation matrix (from the energy value before entry into  $F_1$  to the energy after exit from  $F_2$ ) is no longer given by  $\sum_m c_{nm}d_{ml}$ , but every term in this sum has additionally an unknown

phase factor. No expectation is therefore open to us, except that the mean value of  $e_{nl}\bar{e}_{nl}$  averaged over all these expected phase alterations is equal to  $Z_{nl}$ . A simple calculation confirms that this is the case. We can therefore deduce from one experiment the possible results of another by definite statistical rules. The other experiment itself selects out of the plenitude of all possibilities a quite definite one, and thereby limits the possibilities for all later experiments. Such an interpretation of the equation for the transformation matrix  $S$  or the Schrödinger wave equation is only possible because the sum of solutions is again a solution. In this circumstance we see the deep significance of the linearity of Schrödinger wave equations. On that account they can be understood only as equations for waves in phase space; and on that account we may regard as hopeless every attempt to replace these equations by nonlinear equations, for example in the relativistic case (for more than one electron).

### §3. THE TRANSITION FROM MICRO- TO MACROMECHANICS

It seems to me that the concepts of kinematics and mechanics in quantum theory are sufficiently clarified by the analysis of the words "position of an electron," "velocity," "energy," etc., in the preceding sections that physical understanding of macroscopic processes from the standpoint of quantum mechanics must also be possible. The transition from micro- to macromechanics has already been treated by Schrödinger,\* but I do not believe that Schrödinger's considerations

\* E. Schrödinger, *Naturwiss.*, 14, 664 (1926).

get to the heart of the problem, and this is why: According to Schrödinger, in a high state of excitation a sum of eigenfunctions ought to be able to give a wave packet of limited extent which—through periodic changes in its form—will carry out the periodic motions of the classical “electron.” There is an argument against this outlook: If the wave packet had such properties as ascribed to it by this view, then the radiation sent out by the atom would be representable as a Fourier series in which the frequencies of the higher vibrations were integer multiples of the basic frequency. The frequencies of the spectral lines sent out by the atom are, however, according to quantum mechanics, never integer multiples of the basic frequency—except in the special case of the harmonic oscillator. Thus Schrödinger’s reasoning is only viable for the case of the harmonic oscillator treated by him; in all other cases a wave packet spreads out in the course of time over the whole immediate neighborhood of the atom. The higher the state of excitation of the atom, the slower is that spreading of the wave packet. However, if one waits long enough it happens. The reasoning developed above about the radiation sent out by the atom might at first sight be used against all experiments which look for a direct transition from quantum to classical mechanics at high quantum numbers. For that reason the attempt was made earlier to circumvent such reasoning by referring to the natural radiation broadening of stationary states—certainly wrongly, first of all because this way out is blocked already in the case of the hydrogen atom on account of the weakness of the radiation for high states, and secondly, because the transition from quantum to classical mechanics ought to be understandable without calling on electrodynamics. Bohr\* has already referred many times to these well-known difficulties which stand in the way of a direct connection between quantum and classical theory. We have spelled them out again here so explicitly only because in recent times they seem to be forgotten.

I believe that one can fruitfully formulate the origin of the classical “orbit” in this way: the “orbit” comes into being only when we observe it. For example, let an atom be given in a state of excitation  $n = 1000$ . The dimensions of the orbit in this case are already relatively large so that, in accordance with §1, it is enough to use light of relatively low wavelength to determine the position of the electron. If the position determination is not to be too fuzzy then the Compton recoil will put the atom in some state of excitation between, say, 950 and 1050. Simultaneously, the momentum of the electron can be determined from the Doppler effect with a precision given by (1). One can characterize the experimental finding by a wave-packet, or, better, a probability-amplitude packet, in  $q$ -space of a spread given by the wavelength of the light used, and built up primarily out of eigenfunctions between the 950th and 1050th eigenfunction—and by a corresponding packet in  $p$ -space. Let a new determination of position be made after some time with the same precision. Its result, according to §2, can be predicted only statistically. All

positions count as likely (with calculable probability) which lie within the bounds of the now broadened wavepacket. The situation would be no different in classical theory, for there too the result of the second position measurement would be predictable only statistically because of the uncertainty in the first measurement. Also, the orbits of classical theory would spread out like the wavepacket. However, statistical laws themselves are different in quantum mechanics and in classical theory. The second determination of the position selects a definite “ $q$ ” from the totality of possibilities and limits the options for all subsequent measurements. After the second position determination the results of later measurements can only be calculated when one again ascribes to the electron a “smaller” wavepacket of extension  $\lambda$  (wavelength of the light used in the observation). Thus every position determination reduces the wavepacket back to its original extension  $\lambda$ . The “values” of the quantities  $\mathbf{p}$ ,  $\mathbf{q}$  are known throughout all the experiments with a certain precision. The values of  $\mathbf{p}$  and  $\mathbf{q}$  stay within the precision limits fixed by the classical equations of motion. This can be seen directly from the quantum-mechanical equations,

$$\dot{\mathbf{p}} = -\partial \mathbf{H} / \partial \mathbf{q}; \quad \dot{\mathbf{q}} = \partial \mathbf{H} / \partial \mathbf{p}. \quad (9)$$

However, the orbit, as noted earlier, can be calculated only statistically from the initial conditions, a circumstance that one can consider as a consequence of the fundamental indeterminism of the initial conditions. The statistical laws are different for quantum mechanics and classical theory. This distinction, under appropriate circumstances, can give rise to gross macroscopic differences between classical and quantum theory. Before I discuss an example, I should like to indicate how the transition, discussed above, to classical theory is formulated mathematically for a simple mechanical system, the force-free motion of a particle. In one dimension the equations of motion run

$$\mathbf{H} = \mathbf{p}^2 / 2m; \quad \dot{\mathbf{q}} = \mathbf{p}/m; \quad \dot{\mathbf{p}} = 0. \quad (10)$$

As time can be treated as a parameter (or “ $c$ -number”) when there are no time-dependent external forces, the solution of this equation is

$$\mathbf{q} = \mathbf{p}_0 t / m + \mathbf{q}_0; \quad \mathbf{p} = \mathbf{p}_0, \quad (11)$$

where  $\mathbf{p}_0$  and  $\mathbf{q}_0$  are the momentum and the position at the time  $t = 0$ . At  $t = 0$  the value  $q_0 = q'$  is measured with accuracy  $q_1$ , and  $p_0 = p'$  with accuracy  $p_1$  [see equations (3) to (6)]. In order to draw conclusions from the “values” of  $\mathbf{p}_0$  and  $\mathbf{q}_0$  about the “values” of  $\mathbf{q}$  at the time  $t$ , one must find—according to Dirac and Jordan—that transformation function which transforms all matrices in the 825

representation\* in which  $\mathbf{q}_0$  is diagonal to that representation in which  $\mathbf{q}$  is diagonal. In the matrix scheme in which  $\mathbf{q}_0$  is diagonal,  $\mathbf{p}_0$  can be replaced by the operator  $-i\hbar \hat{c}/\hat{c}q_0$ . According to Dirac [*i.e.* equation (11)] the desired transformation amplitude  $S(q_0, q)$  satisfies the differential equation,

$$\{-i(\hbar t/m)\hat{c}/\hat{c}q_0 + q_0\}S(q_0, q) = qS(q_0, q) \quad (12)$$

or

$$-i(\hbar t/m)\hat{c}S/\hat{c}q_0 = (q_0 - q)S(q_0, q),$$

[with the solution]

$$S(q_0, q) = \text{const } \exp \left[ (im/\hbar t) \int (q - q_0) dq_0 \right]. \quad (13)$$

$S\bar{S}$  is therefore independent of  $q_0$ . In other words, if at the time  $t = 0$  we know  $q_0$  exactly, then at any time  $t > 0$  all values of  $q$  are equally probable; that is, the probability that  $q$  lies in any finite region is quite nil. Physically this is already clear without further investigation. Thus the exact determination of  $q_0$  leads to an infinitely large Compton recoil. The same would naturally apply for an arbitrary mechanical system. If, however,  $q_0$  is known at the time  $t = 0$  only within the range  $q_1$ , and  $p_0$  in the range  $p_1$  [see equation (3)], then

$$S(\eta, q_0) = \text{const } \exp [-(q_0 - q')^2/2q_1^2 + (i/\hbar)p'(q_0 - q')],$$

and the probability [amplitude] function for  $q$  is to be calculated according to the formula,

$$S(\eta, q) = \int S(\eta, q_0)S(q_0, q) dq_0.$$

The result is

$$S(\eta, q) = \text{const } \int \exp \{(im/\hbar t)[q_0(q - tp'/m) - q_0^2/2] - (q' - q_0)^2/2q_1^2\} dq_0. \quad (14)$$

With the abbreviation

$$\beta = t\hbar/mq_1^2, \quad (15)$$

the exponent in (14) becomes

826 \* The word "representation," not employed by Heisenberg himself, is introduced here for clarity. He uses the phrase "matrix scheme." (Translators' note.)

$$- \{q_0^2(1 + i/\beta) - 2q_0[q' + i(q - tp'/m)/\beta] + q'^2\}/2q_1^2.$$

The term with  $q'^2$  can be taken into the constant ( $q$ -independent factor) and the integration gives

$$\begin{aligned} S(\eta, q) &= \text{const exp} \{[q' + i(q - tp'/m)/\beta]^2/[2(1 + i/\beta)q_1^2]\} \\ &= \text{const exp} - \{[(q - tp'/m - i\beta q')^2(1 - i/\beta)]/[2q_1^2(1 + \beta^2)]\}. \end{aligned} \quad (16)$$

It follows that

$$S(\eta, q)\bar{S}(\eta, q) = \text{const exp} - \{[q - tp'/m - q']^2/q_1^2(1 + \beta^2)\}. \quad (17)$$

Thus the electron is located at the time  $t$  at the position  $tp'/m + q'$  with a spread  $q_1(1 + \beta^2)^{1/2}$ . The “wavepacket” or, better, “probability packet” has expanded by the factor  $(1 + \beta^2)^{1/2}$ . According to (15),  $\beta$  is proportional to the time  $t$ , inversely proportional to the mass, as is entirely plausible, and inversely proportional to  $q_1^2$ . Too much precision in  $q_0$  produces great uncertainty in  $p_0$  and thus leads to a large uncertainty in  $q$ . The parameter  $\eta$  which we have brought in above for formal reasons might be left out here in all formulas, as it does not enter the calculation. To illustrate that the difference between the classical and the quantum statistical laws leads to gross macroscopic differences between the results of the two theories, let the reflection of a beam of electrons at a grating be discussed briefly. When the spacing of the rulings is of the order of the de Broglie wavelength of the electrons, then reflection occurs in definite, discrete directions like the reflection of light at a grating. What classical theory gives is grossly and macroscopically different. Nevertheless, from the orbit of an individual electron we can in no way find a contradiction with a classical theory. We might if we could, direct the electron, say, to a definite point on a grating ruling, and then verify that the reflection there violates classical theory. However, when we want to determine the position of the electron so precisely that we can say at what point on a grating ruling it hits, then the electron acquires through this position determination a large velocity, and the de Broglie wavelength of the electron becomes so much shorter that now the reflection really can and will take place approximately as predicted classically, without violating the laws of the quantum theory.

#### §4. DISCUSSION OF A FEW SPECIAL IDEALIZED EXPERIMENTS

According to the physical interpretation of quantum theory aimed at here, the time of transitions or “quantum jumps” must be as concrete and determinable by measurement as, say, energies in stationary states. The spread within which such an instant is specifiable is given according to equation (2) by  $\hbar/\Delta E$ , if  $\Delta E$  designates<sup>827</sup>

the change of energy in a quantum jump.\* We consider, for example, the following experiment. An atom, at time  $t = 0$  in state  $n = 2$  may transit, via radiation, to the ground state,  $n = 1$ . Then, in analogy to equation (7), we can ascribe to the atom the eigenfunction

$$S(t, p) = e^{-\alpha t} \psi(E_2, p) e^{-iE_2 t / \hbar} + (1 - e^{-2\alpha t})^{1/2} \psi(E_1, p) e^{-iE_1 t / \hbar} \quad (18)$$

if we assume that the radiation damping is expressed in a factor of the form  $e^{-\alpha t}$  in the eigenfunction (the real dependence is perhaps not so simple). This atom is sent through an inhomogeneous magnetic field for the determination of its energy level, as is usual in the Stern-Gerlach experiment; yet we also have the inhomogeneous field follow the atomic beam over a long stretch of its path. The consequent acceleration we will measure, say, in this way: we divide the entire stretch that the atomic beam pursues in the magnetic field into short sections, at the end of each of which we determine the deviation of the beam. Depending on the velocity of the atomic beam, the division into intervals of space corresponds for the atom to division into small time intervals  $\Delta t$ . According to §1, equation (2), there corresponds to a time interval  $\Delta t$  a spread in energy of  $\hbar/\Delta t$ . The probability of measuring the definite energy  $E$  can be directly deduced from  $S(p, E)$  and is therefore calculated for the interval from  $n \Delta t$  to  $(n + 1) \Delta t$  as,

$$\int_{n \Delta t}^{(n+1) \Delta t} S(p, t) e^{iEt / \hbar} dt.$$

If the determination "state  $n = 2$ " is made at the time  $(n + 1) \Delta t$ , then for everything later one must ascribe to the atom not the eigenfunction (18) but one which results from (18) when  $t$  is replaced by  $t - (n + 1) \Delta t$ . If, on the contrary, one finds "state  $n = 1$ ," then from that point on one has to attribute to the atom the eigenfunction

$$\psi(E_1, p) e^{-iE_1 t / \hbar}.$$

Thus one will first find for a series of intervals  $\Delta t$  "state  $n = 2$ ," then steadily "state  $n = 1$ ." In order that a distinction between the two states will still be possible,  $\Delta t$  cannot be shrunk below  $\hbar/\Delta E$ . Thus the instant of the transition is determinable within this spread. We imply an experiment of the kind just sketched quite in the spirit of the old formulation of quantum theory founded by Planck Einstein, and Bohr when we speak of the discontinuous change of the energy. As such an experiment can in principle be carried out, an agreement about its outcome must be possible.

In Bohr's basic postulates of quantum theory, the energy of an atom has the

\* Compare W. Pauli, *l.c.*, p. 12.

advantage—just as do the values of the action variables  $J$ —over other determinants of the motion (position of the electron, etc.) in that its numerical value can always be given. This preferred position which the energy has over other quantum-mechanical quantities it owes only to the circumstance that it represents an integral of the equations of motion for closed systems (the energy matrix  $\mathbf{E}$  is a constant). For open systems, in contrast, the energy is not singled out over any other quantum-mechanical quantity. In particular, one will be able to devise experiments in which the phases,  $w$ , of the atom are precisely measurable, but in which then the energy remains in principle undetermined, corresponding to the relation  $\mathbf{Jw} - \mathbf{wJ} = -i\hbar$  or  $J_1 w_1 \sim \hbar$ . Resonance fluorescence is such an experiment. If one irradiates an atom with an eigenfrequency, say  $v_{12} = (E_2 - E_1)/\hbar$ , then the atom vibrates in phase with the external radiation. Then, even in principle, it makes no sense to ask in which state,  $E_1$  or  $E_2$ , the atom is thus vibrating. The phase relation between atom and external radiation may be determined, for example, by the phase relations of large numbers of atoms with one another ([R. W.] Wood's experiments). If one prefers to avoid experiments with radiation then one can also measure the phase relation by carrying out exact position determinations on the electron in the sense of §1 at different times relative to the phase of the light impinging (on many atoms). A “wave function,” say, of the form,

$$S(q, t) = c_2 \psi_2(E_2, q) e^{-i(E_2 t + \beta)/\hbar} + (1 - c_2^2)^{1/2} \psi_1(E_1, q) e^{-iE_1 t/\hbar}, \quad (19)$$

can be ascribed to the individual atom. Here  $c_2$  depends on the strength and  $\beta$  on the phase of the incident light. The probability of a definite position  $q$  is thus

$$\begin{aligned} S(q, t) \bar{S}(q, t) &= c_2^2 \psi_2 \bar{\psi}_2 + (1 - c_2^2) \psi_1 \bar{\psi}_1 \\ &\quad + c_2(1 - c_2^2)^{1/2} \{ \psi_2 \bar{\psi}_1 e^{-i[(E_2 - E_1)t + \beta]/\hbar} + \bar{\psi}_2 \psi_1 e^{+i[(E_2 - E_1)t + \beta]/\hbar} \}. \end{aligned} \quad (20)$$

The periodic term in (20) is experimentally distinguishable from the unperiodic ones, as the determinations of position can be carried out for different phases of the incident light.

In a well-known idealized experiment proposed by Bohr, the atoms of a Stern-Gerlach atomic beam are first excited to a resonance fluorescence at a definite state by incident radiation. After a little way they go through an inhomogeneous magnetic field. The radiation emerging from the atoms can be observed during the whole path, before and after the magnetic field. Before the atoms enter the magnetic field, ordinary resonance radiation takes place; that is, as in dispersion theory, it must be assumed that all atoms send out spherical waves in phase with the incident light. The latter view at first sight contradicts the result that a crude application of the quantum theory of light or the basic rules of quantum theory would give<sup>829</sup> Thus from this view one would conclude that only a few atoms are raised to the

“upper state” through absorption of the light quantum, and that therefore the entire resonance radiation arises from a few intensively radiating centers. It therefore seemed natural in earlier times to say that the concept of light quanta ought to be brought in here only to account for the balance of energy and momentum, and that “in reality” all atoms in the ground state radiate weak and coherent spherical waves. After the atoms have passed the magnetic field, however, there can hardly be any doubt that the atomic beam has divided into two beams of which one corresponds to atoms in the upper state, the other in the lower. If now the atoms in the lower state were to radiate, then we would have a gross violation of the law of conservation of energy. For all the energy of excitation resides in the beam with atoms in the upper state. Still less can there be any doubt that past the magnetic field only the “upper state” beam sends out light, and indeed incoherent light, from the few intensively radiating atoms in the upper state. As Bohr has shown, this idealized experiment makes it especially clear that care is often needed in applying the concept of “stationary state.” The formulation of quantum theory developed here allows a discussion of the Bohr experiment to be carried through without any difficulty. In the external radiation field the phases of the atoms are determined. Therefore it is meaningless to speak of the “energy of the atom.” Also, after the atom has left the radiation field one is not entitled to say that it is in a definite stationary state, insofar as one enquires about the coherence properties of the radiation. However, one can set up an experiment to find out in what state the atom is. The result of this experiment can be stated only statistically. Such an experiment is really performed by the inhomogeneous magnetic field. Beyond the magnetic field the energies of the atoms are well determined and therefore the phases are indeterminate. The resulting radiation is incoherent and comes only from atoms in the upper state. The magnetic field determines the energies and therefore destroys the phase relation. Bohr’s idealized experiment is a very beautiful illustration of the fact that the energy of the atom “in reality” is not a number but a matrix. The conservation law holds for the energy matrix and therefore also for the value of the energy as precisely as it can be measured. In mathematical terms the lifting of the phase relation can be traced out as follows, for example. Let  $Q$  be the coordinates of the center of gravity of the atom, so that one ascribes to the atom, instead of (19), the eigenfunction

$$S(Q, t)S(q, t) = S(Q, q, t). \quad (21)$$

Here  $S(Q, t)$  is a function that, like  $S(\eta, q)$  in (16), differs from zero only in a small neighborhood of a point in  $Q$ -space and propagates with the velocity of the atoms in the direction of the beam. The probability of a relative amplitude  $q$  regardless of  $Q$  is given by the integral of

over  $Q$ —that is, through (20). However, the eigenfunction (21) will change by a calculable amount in a magnetic field and, on account of the different deviation of atoms in the upper and lower state, will have changed beyond the magnetic field into

$$S(Q, q, t) = c_2 S_2(Q, t) \psi_2(E_2, q) e^{-i(E_2 t + \beta)/\hbar} + (1 - c_2^2)^{1/2} S_1(Q, t) \psi_1(E_1, q) e^{-iE_1 t / \hbar}. \quad (22)$$

Here  $S_1(Q, t)$  and  $S_2(Q, t)$  will be functions in  $Q$ -space which differ from zero only in the small neighborhood of a point; but this point is different for  $S_1$  and  $S_2$ . The product  $S_1 S_2$  is therefore zero everywhere. The probability of a relative amplitude  $q$  and a definite value  $Q$  is therefore

$$S(Q, q, t) \bar{S}(Q, q, t) = c_2^2 S_2 \bar{S}_2 \psi_2 \bar{\psi}_2 + (1 - c_2^2) S_1 \bar{S}_1 \psi_1 \bar{\psi}_1. \quad (23)$$

The periodic term of (20) has disappeared and with it the possibility of measuring a phase relation. The statistical result of position determinations will always be the same, whatever the phase of the incident radiation. We may assume that experiments with radiation, the theory of which has not yet been developed, will give the same results about phase relations between atoms and incident radiation.

Finally let us examine the connection\* of equation (2),  $E_1 t_1 \sim \hbar$ , with a complex of problems which Ehrenfest and other investigators have discussed on the basis of Bohr's correspondence principle in two important papers.<sup>†</sup> Ehrenfest and Tolman speak of "weak quantization" when the quantized periodic motion is interrupted through quantum jumps or rather perturbations in intervals of time which can be regarded as not very long compared to the periods of the system. These cases should reveal not only the exact quantum energy values but also energy values which do not differ too much from the quantum values, and these with a smaller and qualitatively predictable *a priori* probability. In quantum mechanics this behavior is to be interpreted in these terms. As the energy is really changed by external perturbations or quantum jumps, every energy measurement, insofar as it is to be unique, must be done in the time between two perturbations. In this way an upper bound is specified for  $t_1$  in the sense of §1. Therefore we measure the energy value  $E_0$  of a quantized state also only within a spread  $E_1 \sim \hbar/t_1$ . Here the question is meaningless even in principle whether the system "really" takes on with the correspondingly lower statistical weight such energy values  $E$  as deviate from  $E_0$ , or whether their experimental realization is to be attributed only to the inaccuracy of the measurement. If  $t_1$  is smaller than the period of the system then it is no longer meaningful to speak of discrete stationary states or discrete energy values.

\* W. Pauli drew my attention to this connection.

<sup>†</sup> P. Ehrenfest and G. Breit (*Zeits. f. Physik*, 9, 207 [1922]) and P. Ehrenfest and R. C. Tolman (*Phys. Rev.*, 24, 287 [1924]). See also the discussion in N. Bohr, *Grundpostulate der Quantentheorie*, l.c. 831

Ehrenfest and Breit in a similar connection draw attention to the following paradox. A rotator, which we will visualize as a gear-wheel, is provided with an attachment which after  $f$  revolutions of the wheel exactly reverses the direction of its rotation. For example, let the gear-wheel mesh with a toothed sliding member which moves on a straight line between two stops. The slider hits a stop after a definite number of rotations and in that way reverses the rotation of the gear-wheel. The true period  $T$  of the system is long in comparison with the rotation period  $t$  of the wheel. The discrete energy levels are densely packed—and the denser the packing, the greater the value of  $T$ . From the standpoint of consistent quantum theory all stationary states have the same statistical weight. Therefore, for sufficiently great  $T$ , practically all energy values occur with equal frequency, in opposition to what would be expected for the rotator. We may sharpen this paradox a little before we treat it from our standpoint. Thus, in order to determine whether the system takes on the discrete energy values belonging to the pure rotator exclusively or particularly often, or whether it assumes with equal probability all possible values (that is, values which correspond to the small energy interval  $h/T$ ), a time  $t_1$  suffices which is small relative to  $T$  (but  $\gg t$ ). In other words, although the long period plays no part at all in such measurements, it appears to express itself in the fact that all possible energy values can occur. We are of the view that, in reality also, such experiments for the determination of the total energy of the system would give all possible energy values with equal probability. The factor responsible for this outcome is not the big period  $T$ , but the sliding member. Even if the system sometimes happens to have an energy identical with the quantized energy value of the simple rotator, it can be modified easily—by external forces acting on the stop—to states which do not correspond to the quantization of the simple rotator.\* The coupled system, rotator-plus-slider, indeed shows a periodicity entirely different from that of the rotator. The solution of the paradox lies rather in a different circumstance. When we want to measure the energy of the rotator alone, we must first break the coupling between the rotator and the slider. In classical theory, when the mass of the slider is sufficiently small, the coupling can be broken without a change in energy; and there, consequently, the energy of the entire system can be equated to that of the rotator (for small slider mass). In quantum mechanics the energy of interaction between slider and rotator is at least of the same order of magnitude as the level spacing of the rotator (even for small slider mass there is a high zero point energy associated with the elastic interaction between rotator and slider). On decoupling, the slider and the rotator individually take their characteristic quantum energies. Consequently, insofar as we are able to measure the energy values of the rotator alone we always find the quantum energy values with experimental accuracy. Even for

\* According to Ehrenfest and Breit this cannot happen, or can happen only rarely, through forces which act on the gear-wheel.

vanishing mass of the slider, however, the energy of the coupled system is different from the energy of the rotator. The energy of the coupled system can take on all possible values (consistent with the  $T$ -quantization) with equal probability.

---

Quantum kinematics and mechanics show far-reaching differences from the ordinary theory. The applicability of classical kinematics and mechanical concepts, however, can be justified neither from our laws of thought nor from experiment. The basis for this conclusion is relation (1),  $p_1 q_1 \sim h$ . As momentum, position, energy, etc. are precisely defined concepts, one does not need to complain that the basic equation (1) contains only qualitative predictions. Moreover, as we can think through qualitatively the experimental consequences of the theory in all simple cases, we will no longer have to look at quantum mechanics as unphysical and abstract.\* Of course we would also like to be able to derive, if possible, the quantitative laws of quantum mechanics directly from the physical foundations—that is, essentially, from relation (1). On this account Jordan has sought to interpret the equation,

$$S(q, q'') = \int S(q, q') S(q', q'') dq',$$

as a probability relation. However, we cannot accept this interpretation (§2). We believe, rather, that for the time being the quantitative laws can be derived out of the physical foundations only by use of the principle of maximum simplicity. If, for example, the X-coordinate of the electron is no longer a “number,” as can be concluded experimentally, according to equation (1), then the simplest assumption conceivable [that does not contradict (1)] is that this X-coordinate is a diagonal term of a matrix whose nondiagonal terms express themselves in an uncertainty or—by transformation—in other ways (see for example §4). The prediction that, say, the velocity in the X-direction is “in reality” not a number but the diagonal term of the matrix, is perhaps no more abstract and no more unvisualizable than the statement that the electric field strengths are “in reality” the time part of an antisymmetric tensor located in space-time. The phrase “in reality” here is as much and as little justified as it is in any mathematical description of natural processes. As soon as one accepts that all quantum-theoretical quantities are “in reality” matrices, the quantitative laws follow without difficulty.

If one assumes that the interpretation of quantum mechanics is already correct

\* Schrödinger describes quantum mechanics as a formal theory of frightening, indeed repulsive, abstractness and lack of visualizability. Certainly one cannot overestimate the value of the mathematical (and to that extent physical) mastery of the quantum-mechanical laws that Schrödinger's theory has made possible. However, as regards questions of physical interpretation and principle, the popular view of wave mechanics, as I see it, has actually deflected us from exactly those roads which were pointed out by the papers of Einstein and de Broglie on the one hand and by the papers of Bohr and by quantum mechanics on the other hand.

in its essential points, it may be permissible to outline briefly its consequences of principle. We have not assumed that quantum theory—in opposition to classical theory—is an essentially statistical theory in the sense that only statistical conclusions can be drawn from precise initial data. The well-known experiments of Geiger and Bothe, for example, speak directly against such an assumption. Rather, in all cases in which relations exist in classical theory between quantities which are really all exactly measurable, the corresponding exact relations also hold in quantum theory (laws of conservation of momentum and energy). But what is wrong in the sharp formulation of the law of causality, “When we know the present precisely, we can predict the future,” is not the conclusion but the assumption. Even in principle we cannot know the present in all detail. For that reason everything observed is a selection from a plenitude of possibilities and a limitation on what is possible in the future. As the statistical character of quantum theory is so closely linked to the inexactness of all perceptions, one might be led to the presumption that behind the perceived statistical world there still hides a “real” world in which causality holds. But such speculations seem to us, to say it explicitly, fruitless and senseless. Physics ought to describe only the correlation of observations. One can express the true state of affairs better in this way: Because all experiments are subject to the laws of quantum mechanics, and therefore to equation (1), it follows that quantum mechanics establishes the final failure of causality.

#### ADDITION IN PROOF

After the conclusion of the foregoing paper, more recent investigations of Bohr have led to a point of view which permits an essential deepening and sharpening of the analysis of quantum-mechanical correlations attempted in this work. In this connection Bohr has brought to my attention that I have overlooked essential points in the course of several discussions in this paper. Above all, the uncertainty in our observation does not arise exclusively from the occurrence of discontinuities, but is tied directly to the demand that we ascribe equal validity to the quite different experiments which show up in the corpuscular theory on one hand, and in the wave theory on the other hand. In the use of an idealized gamma-ray microscope, for example, the necessary divergence of the bundle of rays must be taken into account. This has as one consequence that in the observation of the position of the electron the direction of the Compton recoil is only known with a spread which then leads to relation (1). Furthermore, it is not sufficiently stressed that the simple theory of the Compton effect, strictly speaking, only applies to free electrons. The consequent care needed in employing the uncertainty relation is, as Professor Bohr has explained, essential, among other things, for a comprehensive discussion of the transition from micro- to macromechanics. Finally, the discussion of resonance fluorescence is not entirely correct because the connection between

the phase of the light and that of the electronic motion is not so simple as was assumed. I owe great thanks to Professor Bohr for sharing with me at an early stage the results of these more recent investigations of his—to appear soon in a paper on the conceptual structure of quantum theory—and for discussing them with me.

Copenhagen, Institute for Theoretical Physics of the University.

## I.4 COMPLEMENTARITY

### COMMENTARY OF ROSENFELD (1963)

Complementarity is no system, no doctrine with ready-made precepts. There is no *via regia* to it; no formal definition of it can even be found in Bohr's writings, and this worries many people. The French are shocked by this breach of the Cartesian rules; they blame Bohr for indulging in "clair-obscur" and shrouding himself in "les brumes du Nord." The Germans in their thoroughness have been at work distinguishing several forms of complementarity and studying, in hundreds of pages, their relations to Kant. Pragmatic Americans have dissected complementarity with the scalpel of symbolic logic and undertaken to define this gentle art of the correct use of words without using any words at all. Bohr was content to teach by example. He often evoked the thinkers of the past who had intuitively recognized dialectical aspects of existence and endeavored to give them poetical or philosophical expression; our only advantage over these great men, he would observe, is that in physics we have been presented with such a simple and clear case of complementarity that we are able to study it in detail and thus arrive at a precise formulation of a logical relationship of universal scope. The nature of this relation he regarded as sufficiently illustrated by his analyses of the limits of validity of classical physical concepts.

### COMMENTARY OF ROSENFELD (1971A; CONTINUED FROM §I. 3 ABOVE)

As to Bohr's "forthcoming" publication, more than a year elapsed before it appeared in print: it was a much furbished version of a lecture he delivered, shortly after the events just retraced, at a physicists' conference in Como. He had the greatest misgivings about presenting his conception of complementarity to the community of physicists in a state which he judged immature; but he yielded to the advice of his more practically-minded brother Harald. Upon the latter's urging, he even consented to write up a brief account, that could be promptly published in *Nature*, as a letter to the editor: but this letter never reached its destination. With the help of Klein, he actually managed to complete it just on the night of his reluctant departure for Como. However, when Klein came up to the Institute the next morning, and enquired whether the letter had been sent off, he learned that there had been a double hitch, of a kind to delight Freudians. Bohr had missed the night train, because he could not find his passport (which lay on his desk); he had departed by the next train, taking with him the famous letter.

## *The Quantum Theory of Dispersion.*

By P. A. M. DIRAC, St. John's College, Cambridge ; Institute for Theoretical Physics, Göttingen.

(Communicated by R. H. Fowler, F.R.S.—Received April 4, 1927.)

### *§ 1. Introduction and Summary.*

The new quantum mechanics could at first be used to answer questions concerning radiation only through analogies with the classical theory. In Heisenberg's original matrix theory, for instance, it is assumed that the matrix elements of the polarisation of an atom determine the emission and absorption of radiation analogously to the Fourier components in the classical theory. In more recent theories\* a certain expression for the electric density obtained from the quantum mechanics is used to determine the emitted radiation by the same formulæ as in the classical theory. These methods give satisfactory results in many cases, but cannot even be applied to problems where the classical analogies are obscure or non-existent, such as resonance radiation and the breadths of spectral lines.

A theory of radiation has been given by the author which rests on a more definite basis.† It appears that one can treat a field of radiation as a dynamical system, whose interaction with an ordinary atomic system may be described by a Hamiltonian function. The dynamical variables specifying the field are the energies and phases of its various harmonic components, each of which

\* E. Schrödinger, 'Ann. d. Physik,' vol. 81, p. 109 (1926) ; W. Gordon, 'Z. f. Physik,' vol. 40, p. 117 (1926) ; O. Klein, 'Z. f. Physik,' vol. 41, p. 407 (1927).

† 'Roy. Soc. Proc.,' A, vol. 114, p. 243 (1927). This is referred to later by *loc. cit.*

is effectively a simple harmonic oscillator. One must, of course, in the quantum theory take these variables to be q-numbers satisfying the proper quantum conditions. One finds then that the Hamiltonian for the interaction of the field with an atom is of the same form as that for the interaction of an assembly of light-quanta with the atom. There is thus a complete formal reconciliation between the wave and light-quantum points of view.

In applying the theory to the practical working out of radiation problems one must use a perturbation method, as one cannot solve the Schrödinger equation directly. One can assume that the term ( $V$  say) in the Hamiltonian due to the interaction of the radiation and the atom is small compared with that representing their proper energy, and then use  $V$  as the perturbing energy. Physically the assumption is that the mean life time of the atom in any state is large compared with its periods of vibration. In the present paper we shall apply the theory to determine the radiation scattered by the atom, considering also the case when the frequency of the incident radiation coincides with that of a spectral line of the atom. The method used will be that in which one finds a solution of the Schrödinger equation that satisfies certain initial conditions, corresponding to a given initial state for the atom and field. In general terms it may be described as follows :—

If  $V_{mn}$  are the matrix elements of the perturbing energy  $V$ , where each suffix  $m$  or  $n$  refers to a stationary state of the whole system of atom plus field the stationary state of the atom being specified by its action variables,  $J$  say, and that of the field by a given distribution of energy among its harmonic components, or by a given distribution of light-quanta), then each  $V_{mn}$  gives rise to transitions from state  $n$  to state  $m^*$ ; more accurately, it causes the eigenfunction representing state  $m$  to grow if that representing state  $n$  is already excited, the general formula for the rate of change of the amplitude  $a_m$  of an eigenfunction being†

$$ih/2\pi \cdot \dot{a}_m = \sum_n V_{mn} a_n = \sum_n v_{mn} a_n e^{2\pi i (W_m - W_n) t/\hbar}, \quad (1)$$

where  $v_{mn}$  is the constant amplitude of the matrix element  $V_{mn}$ , and  $W_m$  is the

\* In *loc. cit.*, § 6, it was in error assumed that  $V_{mn}$  caused transitions from state  $m$  to state  $n$ , and consequently the information there obtained about an absorption (or emission) process in terms of the number of light-quanta existing before the process should really apply to an emission (or absorption) process in terms of the number of light-quanta in existence after the process. This change, of course, does not affect the results (namely the proof of Einstein's laws) which can depend on  $|V_{mn}|^2 = |V_{nm}|^2$ .

† *Loc. cit.*, equation (4). In the present paper  $\hbar$  is taken to mean just Planck's constant [instead of  $(2\pi)^{-1}$  times this quantity as in *loc. cit.*] which is preferable when one has to deal much with quanta  $\hbar\nu$  of radiation.

total proper energy of the state  $m$ . To solve these equations one obtains a first approximation by substituting for the  $a$ 's on the right-hand side their initial values, a second approximation by substituting for these  $a$ 's their values given by the first approximation, and so on. One or two such approximations will usually be sufficient to give a solution that is fairly accurate for times that are small compared with the life time, but may all the same be large compared with the periods of the atom. From the first approximation, namely,

$$a_m = a_{mo} + \sum_n v_{mn} a_{no} (1 - e^{2\pi i (W_m - W_n) t/\hbar}) / (W_m - W_n), \quad (2)$$

where  $a_{no}$  denotes the initial value of  $a_n$ , one sees readily that when two states  $m$  and  $n$  have appreciably different proper energies, the amplitude  $a_m$  gets changed only by a small extent, varying periodically with the time, on account of transitions from state  $n$ . Only when two states,  $m$  and  $m'$  say, have the same energy does the amplitude  $a_m$  of one of them grow continually at the expense of that of the other, as is necessary for physically recognisable transitions to occur, and the rate of growth is then proportional to  $v_{mm'}$ .

The interaction term of the Hamiltonian function obtained in *loc. cit.* [equation (30)] does not give rise to any direct scattering processes, in which a light-quantum jumps from one state to another of the same frequency but different direction of motion (*i.e.*, the corresponding matrix element  $v_{mm'} = 0$ ). All the same, radiation that has apparently been scattered can appear by a double process in which a third state,  $n$  say, with different proper energy from  $m$  and  $m'$ , plays a part. If initially all the  $a$ 's vanish except  $a_{m'}$ , then  $a_n$  gets excited on account of transitions from state  $m'$  by an amount proportional to  $v_{nm'}$ , and although it must itself always remain small, a calculation shows that it will cause  $a_m$  to grow continually with the time at a rate proportional to  $v_{mn}v_{nm'}$ . The scattered radiation thus appears as the result of the two processes  $m' \rightarrow n$  and  $n \rightarrow m$ , one of which must be an absorption and the other an emission, in neither of which is the total proper energy even approximately conserved.

The more accurate expression for the interaction energy obtained in § 3 of the present paper does give rise to direct scattering processes, whose effect is of the same order of magnitude as that of the double processes, and must be added to it. The sum of the two will be found to give just Kramers' and Heisenberg's dispersion formula\* when the incident frequency does not coincide with that of an absorption or emission line of the atom. If, however, the incident frequency coincides with that of, say, an absorption line, one of the

\* Kramers and Heisenberg, 'Z. f. Physik,' vol. 31, p. 681 (1925).

terms in the Kramers-Heisenberg formula becomes infinite. The present theory shows that in this case the scattered radiation consists of two parts, of which the amount of one increases proportionally to the time since the interaction commenced, and that of the other proportionally to the square of this time. The first part arises from those terms in the Kramers-Heisenberg formula that remain finite, with perhaps a contribution from the infinite term, while the second, which is much larger, is just what one would get from transitions of the atom to the upper state and down again governed by Einstein's laws of absorption and emission.

A difficulty that appears in the present treatment of radiation problems should be here pointed out. If one tries to calculate, for instance, the total probability of a light-quantum having been emitted by a given time, one obtains as result a sum or integral with respect to the frequency of the emitted light-quantum that does not converge in the high frequencies. This difficulty is not due to any fundamental mistake in the theory, but comes from the fact that the atom has, for the purpose of its interaction with the field, been counted simply as a varying electric dipole, and the field produced by a dipole, when resolved into its Fourier components, has an infinite amount of energy in the short wave-lengths, owing to the infinite field in its immediate neighbourhood. If one does not make the approximation of regarding the atom as a dipole, but uses the exact expression for the interaction energy, then the fact that the singularity in the field is of a lower order of magnitude and remains constant is sufficient to make the series or integral converge. The exact interaction energy is too complicated to be used as a basis for radiation theory at present, and we shall here use only the dipole energy, which will mean that divergent series are always liable to appear in the calculation. The best method to adopt under such circumstances is first to work out the general theory of any effect using arbitrary coefficients  $v_{mn}$ , and then to substitute for these coefficients in the final result their values given by the dipole interaction energy. If one then finds that the series all converge, one can assume that the result is a correct first approximation; if, however, any of them do not converge, one must conclude that a dipole theory is inadequate for the treatment of that particular effect. We shall find that for the phenomena of dispersion and resonance radiation dealt with in the present paper, there are no divergent series in the first approximation, so that the dipole theory is sufficient. If, however, one tries to calculate the breadth of a spectral line, one meets with a divergent series, so that a dipole theory of the atom is presumably inadequate for the correct treatment of this question.

§ 2. *Preliminary Formulae.*

We consider the electromagnetic field to be resolved into its components of plane, plane-polarised, progressive waves, each component  $r$  having a definite frequency, direction of motion and state of polarisation, and being associated with a certain type of light-quanta. (To save writing we shall in future suppose the words "direction of motion" applied to a light-quantum or a component of the field to imply also its state of polarisation, and a sum or integral taken over all directions of motion to imply also the summation over both states of polarisation for each direction of motion. This is convenient because the two variables, direction of motion and state of polarisation, are always treated mathematically in the same way.) For an electromagnetic field of infinite extent there will be a continuous three-dimensional range of these components. As this would be inconvenient to deal with mathematically, we suppose it to be replaced by a large number of discrete components. If there are  $\sigma_r$  components per unit solid angle of direction of motion per unit frequency range, we can keep  $\sigma_r$ , an arbitrary function of the frequency and direction of motion of the component  $r$ , provided it is large and reasonably continuous, and shall find that it always cancels from the final results of a calculation, which fact appears to justify our replacement of the continuous range by the discrete set.

We can express  $\sigma_r$  in the form  $\sigma_r = (\Delta\nu_r \Delta\omega_r)^{-1}$ , where  $\Delta\nu_r$  can be regarded as the frequency interval between successive components in the neighbourhood of the component  $r$ , and  $\Delta\omega_r$  is in the same way the solid angle of direction of motion to be associated with this component. The quantities  $\Delta\nu_r$ ,  $\Delta\omega_r$  enable one to pass directly from sums to integrals. Thus if  $f_r$  is any function of the frequency and direction of motion of the component  $r$  that varies only slightly from one component to a neighbouring one, the sum of  $f_r \Delta\nu_r$  for all components having a specified direction of motion is

$$\Sigma_\nu f_r \Delta\nu_r = \int f_r d\nu_r, \quad (3)$$

and the sum of  $f_r \Delta\omega_r$  for all components having a specified frequency is

$$\Sigma_\omega f_r \Delta\omega_r = \int f_r d\omega_r. \quad (3')$$

Also the sum of  $f_r (\sigma_r)^{-1}$  for all components is

$$\Sigma f_r (\sigma_r)^{-1} = \Sigma f_r \Delta\nu_r \Delta\omega_r = \int f_r d\nu_r d\omega_r. \quad (3'')$$

If the number\*  $N_s$  of quanta of energy of the component  $s$  varies only slightly from one component to a neighbouring one, one can give a meaning to the intensity of the radiation per unit frequency range. By supposing the discreteness in the number of components to arise from the radiation being confined in an enclosure (which would imply stationary waves and a special function  $\sigma_s$ ) one obtains† for the rate of flow of energy per unit area per unit solid angle per unit frequency range

$$I_{\nu} = N_s h \nu_s^3 / c^2, \quad (4)$$

a result which may be taken to hold generally for arbitrary  $\sigma_s$  and progressive waves.‡ If only those components with a specified direction of motion are excited, we have instead that the rate of flow of energy per unit area per unit frequency range is

$$I_{\nu} = N_s h \nu_s^3 / c^2 \cdot \Delta \omega_s; \quad (5)$$

while if only a single component  $s$  is excited, we have that the rate of flow of energy per unit area is

$$I = N_s h \nu_s^3 / c^2 \cdot \Delta \omega_s \Delta \nu_s = N_s h \nu_s^3 / c^2 \sigma_s. \quad (6)$$

In this last case the amplitude of the electric force has the value  $E$  given by

$$E^2 = 8\pi I / c = 8\pi N_s h \nu_s^3 / c^3 \sigma_s, \quad (7)$$

and the amplitude  $a$  of the magnetic vector potential, when chosen so that the electric potential is zero, is

$$a = cE / 2\pi \nu_s = 2(h\nu_s / 2\pi c \sigma_s)^{\frac{1}{2}} N_s^{\frac{1}{2}}. \quad (8)$$

### § 3. The Hamiltonian Function.

We shall now determine the Hamiltonian function that describes the interaction of the field with an atom more accurately than in *loc. cit.* We consider the atom to consist of a single electron moving in an electrostatic field of potential  $\phi$ . According to the classical theory its relativity Hamiltonian equation when undisturbed is

$$p_x^2 + p_y^2 + p_z^2 - (W + e\phi)^2 / c^2 + m^2 c^2 = 0,$$

so that its Hamiltonian function is

$$H = W = c \{m^2 c^2 + p_x^2 + p_y^2 + p_z^2\}^{\frac{1}{2}} - e\phi. \quad (9)$$

\* The rule given in *loc. cit.* that symbols representing c-number values for q-number variables should be primed need not always be observed if no confusion thus arises, as in the present case.

† *Loc. cit.*, § 6, equation (28).

‡ This is justified by the fact that one can obtain the result by an alternative method that does not require a finite enclosure, namely by using a quantum-mechanical argument similar to that of *loc. cit.* (lower part of p. 259), applied to the case of discrete momentum values.

If now there is a perturbing field of radiation, given by the magnetic vector potential  $\kappa_x, \kappa_y, \kappa_z$  chosen so that the electric scalar potential is zero, the Hamiltonian equation for the perturbed system will be

$$\left(p_x + \frac{e}{c}\kappa_x\right)^2 + \left(p_y + \frac{e}{c}\kappa_y\right)^2 + \left(p_z + \frac{e}{c}\kappa_z\right)^2 - \frac{(W + e\phi)^2}{c^2} + m^2c^2 = 0,$$

which gives for the Hamiltonian function

$$\begin{aligned} H = W = c & \left\{ m^2c^2 + \left(p_x + \frac{e}{c}\kappa_x\right)^2 + \left(p_y + \frac{e}{c}\kappa_y\right)^2 \left(p_z + \frac{e}{c}\kappa_z\right)^2 \right\}^{\frac{1}{2}} - e\phi \\ & = c \{ [m^2c^2 + p_x^2 + p_y^2 + p_z^2] + [2e/c \cdot (p_x\kappa_x + p_y\kappa_y + p_z\kappa_z) \\ & \quad + e^2/c^2 \cdot (\kappa_x^2 + \kappa_y^2 + \kappa_z^2)] \}^{\frac{1}{2}} - e\phi. \end{aligned}$$

By expanding the square root, counting the second term in square brackets [ ] as small, and then neglecting relativity corrections for this term, one finds approximately

$$\begin{aligned} H & = c[m^2c^2 + p_x^2 + p_y^2 + p_z^2]^{\frac{1}{2}} - e\phi + e/c \cdot (\dot{x}\kappa_x + \dot{y}\kappa_y + \dot{z}\kappa_z) \\ & \quad + e^2/2mc^2 \cdot (\kappa_x^2 + \kappa_y^2 + \kappa_z^2) \\ & = H_0 + e/c \cdot (\dot{x}\kappa_x + \dot{y}\kappa_y + \dot{z}\kappa_z) + e^2/2mc^2 \cdot (\kappa_x^2 + \kappa_y^2 + \kappa_z^2), \end{aligned} \quad (10)$$

where  $H_0$  is the Hamiltonian for the unperturbed system given by (9). When one counts the radiation field as a dynamical system, one must add on its proper energy  $\Sigma N_r h\nu_r$  to the Hamiltonian (10).

According to the classical theory, the magnetic vector potential for any component  $r$  of the radiation is

$$\kappa_r = a_r \cos 2\pi\theta_r/h = 2(h\nu_r/2\pi c\sigma_r)^{\frac{1}{2}} N_r^{\frac{1}{2}} \cos 2\pi\theta_r/h \quad (11)$$

from (8), where  $\theta_r$  increases uniformly with the time such that  $\dot{\theta}_r = h\nu_r$ , and is the variable that must be taken to be the canonical conjugate of  $N_r$  when the radiation field is treated as a dynamical system. The direction of this vector potential is that of the electric vector of the component of radiation. Hence the total value of the component of the vector potential in any direction, say that of the  $x$ -axis, is

$$\kappa_x = \sum_r \kappa_r \cos \alpha_{xr} = 2(h/2\pi c)^{\frac{1}{2}} \sum_r \cos \alpha_{xr} (\nu_r/\sigma_r)^{\frac{1}{2}} N_r^{\frac{1}{2}} \cos 2\pi\theta_r/h, \quad (12)$$

where  $\alpha_{xr}$  is the angle between the electric vector of the component  $r$  and the  $x$ -axis. In the quantum theory, where the variables  $N_r, \theta_r$  are q-numbers, the expression  $2N_r^{\frac{1}{2}} \cos 2\pi\theta_r/h$  must be replaced by the real q-number  $N_r^{\frac{1}{2}} e^{2\pi i\theta_r/h} + (N_r + 1)^{\frac{1}{2}} e^{-2\pi i\theta_r/h}$ . With this change one can take over the

Hamiltonian (10) into the quantum theory, which gives, when one includes the term  $\Sigma N_r h \nu_r$ ,

$$\begin{aligned} H = H_0 + \Sigma N_r h \nu_r + e h^{\frac{1}{2}} / (2\pi)^{\frac{1}{2}} c^{\frac{3}{2}} \cdot \Sigma_r \omega_r (\nu_r / \sigma_r)^{\frac{1}{2}} [N_r^{\frac{1}{2}} e^{2\pi i \theta_r / \hbar} + (N_r + 1)^{\frac{1}{2}} e^{-2\pi i \theta_r / \hbar}] \\ + e^2 h / 4\pi m c^3 \cdot \Sigma_{r,s} \cos \alpha_{rs} (\nu_r \nu_s / \sigma_r \sigma_s)^{\frac{1}{2}} [N_r^{\frac{1}{2}} e^{2\pi i \theta_r / \hbar} + (N_r + 1)^{\frac{1}{2}} e^{-2\pi i \theta_r / \hbar}] \\ \times [N_s^{\frac{1}{2}} e^{2\pi i \theta_s / \hbar} + (N_s + 1)^{\frac{1}{2}} e^{-2\pi i \theta_s / \hbar}] \quad (13) \end{aligned}$$

where  $x_r$  denotes the component of the vector  $(x, y, z)$  in the direction of the electric vector of the component  $r$ , *i.e.*,

$$x_r = x \cos \alpha_{xr} + y \cos \alpha_{yr} + z \cos \alpha_{zr},$$

and  $\alpha_{rs}$  denotes the angle between the electric vectors of the components  $r$  and  $s$ , *i.e.*

$$\cos \alpha_{rs} = \cos \alpha_{xr} \cos \alpha_{xs} + \cos \alpha_{yr} \cos \alpha_{ys} + \cos \alpha_{zr} \cos \alpha_{zs}.$$

The terms in the first line of (13) are just those obtained in *loc. cit.*, equation (30), and give rise only to emission and absorption processes. The remaining terms (*i.e.*, those in the double summation) were neglected in *loc. cit.*. These terms may be divided into three sets :—

(i) Those terms that are independent of the  $\theta$ 's, which can be added to the proper energy  $H_0 + \Sigma N_r h \nu_r$ . The sum of all such terms, which can arise only when  $r = s$ , is

$$\begin{aligned} e^2 h / 4\pi m c^3 \cdot \Sigma_r \nu_r / \sigma_r \cdot [N_r^{\frac{1}{2}} e^{2\pi i \theta_r / \hbar} (N_r + 1)^{\frac{1}{2}} e^{-2\pi i \theta_r / \hbar} \\ + (N_r + 1)^{\frac{1}{2}} e^{-2\pi i \theta_r / \hbar} N_r^{\frac{1}{2}} e^{2\pi i \theta_r / \hbar}] \\ = e^2 h / 4\pi m c^3 \cdot \Sigma_r \nu_r / \sigma_r \cdot (2N_r + 1). \end{aligned}$$

The terms  $e^2 h / 4\pi m c^3 \cdot \Sigma \nu_r / \sigma_r \cdot 2N_r$  are negligible compared with  $\Sigma N_r h \nu_r$ , owing to the very large quantity  $\sigma_r$  in the denominator, while the terms  $e^2 h / 4\pi m c^3 \cdot \Sigma \nu_r / \sigma_r$  may be ignored since they do not involve any of the dynamical variables, in spite of the fact that the sum  $\Sigma \nu_r / \sigma_r$ , equal to  $\int \nu_r d\nu_r d\omega_r$  from (3''), does not converge for the high frequencies.

(ii) The terms containing a factor of the form  $e^{2\pi i(\theta_r - \theta_s) / \hbar}$  ( $r \neq s$ ), whose sum is

$$\begin{aligned} e^2 h / 4\pi m c^3 \Sigma_r \Sigma_{s \neq r} \cos \alpha_{rs} (\nu_r \nu_s / \sigma_r \sigma_s)^{\frac{1}{2}} [N_r^{\frac{1}{2}} (N_s + 1)^{\frac{1}{2}} e^{2\pi i(\theta_r - \theta_s) / \hbar} \\ + (N_r + 1)^{\frac{1}{2}} N_s^{\frac{1}{2}} e^{-2\pi i(\theta_r - \theta_s) / \hbar}] \\ = e^2 h / 2\pi m c^3 \Sigma_r \Sigma_{s \neq r} \cos \alpha_{rs} (\nu_r \nu_s / \sigma_r \sigma_s)^{\frac{1}{2}} N_r^{\frac{1}{2}} (N_s + 1)^{\frac{1}{2}} e^{2\pi i(\theta_r - \theta_s) / \hbar}. \quad (14) \end{aligned}$$

These terms, which are the only important ones in the three sets, give rise to transitions in which a light-quantum jumps directly from a state  $s$  to a state  $r$ .

Such transitions may be called true scattering processes, to distinguish them from the double scattering processes described in § 1.

(iii) The remaining terms, each of which involves a factor of one or other of the forms  $e^{\pm 4\pi i \theta_r/\hbar}$ ,  $e^{\pm 2\pi i (\theta_r + \theta_s)/\hbar}$ . These terms correspond to processes in which two light-quanta are emitted or absorbed simultaneously, and cannot arise in a light-quantum theory in which there are no forces between the light quanta. The effects of these terms will be found to be negligible, so that the disagreement with the light-quantum theory is not serious.

#### § 4. Discussion of the Emission and True Scattering Processes.

We shall consider now the simple emission processes, in order to discuss the divergent integral that arises in this question. Suppose a light-quantum to be emitted in state  $r$ , with a simultaneous jump of the atom from the state  $J = J'$  to the state  $J = J''$ . If we label the final state of the whole system of atom plus field  $m$  and the initial state  $k$ , the value at time  $t$  of the amplitude  $a_m$  of the eigenfunction of the final state will be in the first approximation

$$a_m = v_{mk} (1 - e^{2\pi i (W_m - W_k)t/\hbar}) / (W_m - W_k), \quad (15)$$

obtained by putting  $a_{k_0} = 1$ ,  $a_{n_0} = 0$  ( $n \neq k$ ) in equation (2). The only term in the Hamiltonian (13) that can contribute anything to the matrix element  $v_{mk}$  is the one involving  $e^{2\pi i \theta_r/\hbar}$ , whose  $(J'', N_1', N_2' \dots N_r' + 1 \dots; J', N_1, N_2 \dots N_r \dots)$  matrix element is  $eh^{\frac{1}{2}}/(2\pi)^{\frac{1}{2}} c^{\frac{3}{2}} \cdot \dot{x}_r(J''J') (\nu_r/\sigma_r)^{\frac{1}{2}} (N_r' + 1)^{\frac{1}{2}}$ , where  $\dot{x}_r(J''J')$  is the ordinary  $(J''J')$  matrix element of  $\dot{x}_r$ . If there is no incident radiation we must take all the  $N$ 's zero, which gives

$$v_{mk} = eh^{\frac{1}{2}} / (2\pi)^{\frac{1}{2}} c^{\frac{3}{2}} \cdot \dot{x}_r(J''J') (\nu_r/\sigma_r)^{\frac{1}{2}},$$

and also

$$W_k = H_o(J') \quad W_m = H_o(J'') + \hbar \nu_r.$$

Thus

$$W_m - W_k = H_o(J'') + \hbar \nu_r - H_o(J') = \hbar [\nu_r - \nu(J' J'')]$$

where  $\nu(J' J'') = [H_o(J') - H_o(J'')]/\hbar$  is the transition frequency between states  $J'$  and  $J''$ , if one assumes  $J'$  to be the higher one. Hence from (15)

$$|a_m|^2 = \frac{e^2}{\pi \hbar c^3} |\dot{x}_r(J''J')|^2 \frac{\nu_r}{\sigma_r} \frac{1 - \cos 2\pi [\nu_r - \nu(J' J'')] t}{[\nu_r - \nu(J' J'')]^2}.$$

To obtain the total probability of any light-quantum being emitted within the solid angle  $\delta\omega$  about the direction of motion of a given light-quantum  $r$  with this jump of the atom, we must multiply  $|a_m|^2$  by  $\delta\omega/\Delta\omega_r$  and sum for all frequencies. This gives, with the help of (3)

$$\delta\omega \sum_\nu \frac{|a_m|^2}{\Delta\omega_r} = \delta\omega \frac{e^2}{\pi \hbar c^3} |\dot{x}_r(J''J')|^2 \int_0^\infty \nu_r d\nu_r \frac{1 - \cos 2\pi [\nu_r - \nu(J' J'')] t}{[\nu_r - \nu(J' J'')]^2}. \quad (16)$$

The integral does not converge for the high frequencies. This is due, as mentioned in § 1, to the non-legitimacy of taking only the dipole action of the atom into account, which is what one does when one substitutes for the magnetic potential in (10) its value given by (12), which is its value at some fixed point such as the nucleus instead of its value where the electron is momentarily situated. To obtain the interaction energy exactly, one should put  $\cos 2\pi [\theta_r/h - v_r \xi_r/c]$  instead of  $\cos 2\pi 0r/h$  in (11), where  $\xi_r$  is the component of the vector  $(x, y, z)$  in the direction of motion of the component  $r$  of radiation. This will make no appreciable change for low frequencies  $v_r$ , but will cause a new factor  $\cos 2\pi v_r \xi_r/c$  or  $\sin 2\pi v_r \xi_r/c$ , whose matrix elements tend to zero as  $v_r$  tends to infinity, to appear in the coefficients of (13). This will presumably cause the integral in (16) to converge when corrected, as its divergence when uncorrected is only logarithmic.

Assuming that the integrand in (16) has been suitably modified in the high frequencies, one sees that for values of  $t$  large compared with the periods of the atom (but small compared with the life time in order that the approximations may be valid) practically the whole of the integral is contributed by values of  $v_r$  close to  $v(J' J'')$ , which means physically that only radiation close to a transition frequency can be spontaneously emitted. One finds readily for the total probability of the emission, by performing the integration,

$$\delta \omega e^2/\pi hc^3 \cdot |\dot{x}_r(J' J'')|^2 \cdot 2\pi^2 t v(J' J''),$$

which leads to the correct value for Einstein's A coefficient per unit solid angle, namely,

$$2\pi e^2/hc^3 \cdot |\dot{x}_r(J' J'')|^2 v(J' J'') = 8\pi^3 e^2/hc^3 \cdot |\dot{x}_r(J' J'')|^2 v^3(J' J'').$$

We shall now determine the rate at which true scattering processes occur, caused by the terms (14) in the Hamiltonian. We see at once that the frequency of occurrence of these processes is independent of the nature of the atom, and is thus the same for a bound as for a free electron. The true scattering is the only kind of scattering that can occur for a free electron, so that we should expect the terms (14) to lead to the correct formula for the scattering of radiation by a free electron, with neglect of relativity mechanics and thus of the Compton effect.

Suppose that initially the atom is in the state  $J'$  and all the  $N$ 's vanish except one of them,  $N_s$  say, which has the value  $N'_s$ . We label this state for the whole system by  $k$ , and the state for which  $J = J'$  and  $N_s = N'_s - 1$ ,  $N_r = 1$  with

all the other N's zero by  $m$ . In the first approximation  $a_m$  is again given by (15), where we now have

$$v_{mk} = e^2 \hbar / 2\pi m c^3 \cdot \cos \alpha_{rs} (\nu_r \nu_s / \sigma_r \sigma_s)^{1/2} N_s'^{1/2}, \quad (17)$$

$$W_k = H_0(J') + N_s' h \nu_s, \quad W_m = H_0(J') + (N_s' - 1) h \nu_s + h \nu_r. \quad (18)$$

Thus

$$W_m - W_k = h (\nu_r - \nu_s), \quad (19)$$

and hence

$$|a_m|^2 = \frac{e^4}{2\pi^2 m^2 c^6} \cos^2 \alpha_{rs} \frac{\nu_r \nu_s}{\sigma_r \sigma_s} N_s' \frac{1 - \cos 2\pi(\nu_r - \nu_s)t}{(\nu_r - \nu_s)^2}.$$

To obtain the total probability of a scattered light-quantum being in the solid angle  $\delta\omega$  we must, as before, multiply  $|a_m|^2$  by  $\delta\omega/\Delta\omega_r$  and sum for all frequencies  $\nu_r$ , which gives\*

$$\delta\omega \sum_r \frac{|a_m|^2}{\Delta\omega_r} = \delta\omega \frac{e^4}{2\pi^2 m^2 c^6} \cos^2 \alpha_{rs} \frac{\nu_s}{\sigma_s} N_s' \int \nu_r d\nu_r \frac{1 - \cos 2\pi(\nu_r - \nu_s)t}{(\nu_r - \nu_s)^2}. \quad (20)$$

We again obtain a divergent integral, of the same form as before, which we may assume becomes convergent in the more exact theory. We now have that practically the whole of the integral is contributed by values of  $\nu_r$  close to  $\nu_s$  and the total probability for the scattering process is

$$\delta\omega \frac{e^4}{2\pi^2 m^2 c^6} \cos^2 \alpha_{rs} \frac{\nu_s}{\sigma_s} N_s' \cdot 2\pi^2 t \nu_s = \delta\omega \frac{e^4}{h m^2 c^4 \nu_s} \cos^2 \alpha_{rs} \cdot t I$$

from (6), where  $I$  is the rate of flow of incident energy per unit area. The rate of emission of scattered energy per unit solid angle is thus

$$e^4 / m^2 c^4 \cdot \cos^2 \alpha_{rs} I,$$

where  $\alpha_{rs}$  is the angle between the electric vectors of the incident and scattered radiation, which is the correct classical formula.

\* The reason why there is a small probability for the scattered frequency  $\nu_r$  differing by a finite amount from the incident frequency  $\nu_s$  is because we are considering the scattered radiation, after the scattering process has been acting for only a finite time  $t$ , resolved into its Fourier components. One sees from the formula (20) that as the time  $t$  gets greater, the scattered radiation gets more and more nearly monochromatic with the frequency  $\nu_s$ . If one obtained a periodic solution of the Schrödinger equation corresponding to permanent physical conditions, one would then find that the scattered frequency was exactly equal to the incident frequency.

## § 5. Theory of Dispersion.

We shall now work out the second approximation to the solution of equations (1), taking the case when the system is initially in the state  $k$ , so that the first approximation, given by (2) with  $a_{no} = \delta_{nk}$ , reduces to

$$a_m = \delta_{mk} + v_{mk} (1 - e^{2\pi i (W_m - W_k)t/\hbar}) / (W_m - W_k).$$

When one substitutes these values for the  $a_n$ 's in the right-hand side of (1), one obtains

$$\begin{aligned} i\hbar/2\pi \cdot \dot{a}_m &= v_{mk} e^{2\pi i (W_m - W_k)t/\hbar} \\ &\quad + \sum_n v_{mn} v_{nk} (1 - e^{2\pi i (W_n - W_k)t/\hbar}) e^{2\pi i (W_m - W_n)t/\hbar} / (W_n - W_k) \\ &= \left( v_{mk} - \sum_n \frac{v_{mn} v_{nk}}{W_n - W_k} \right) e^{2\pi i (W_m - W_k)t/\hbar} + \sum_n \frac{v_{mn} v_{nk}}{W_n - W_k} e^{2\pi i (W_m - W_n)t/\hbar}, \end{aligned}$$

and hence when  $m \neq k$

$$\begin{aligned} a_m &= \left( v_{mk} - \sum_n \frac{v_{mn} v_{nk}}{W_n - W_k} \right) \frac{1 - e^{2\pi i (W_m - W_k)t/\hbar}}{W_m - W_k} \\ &\quad + \sum_n \frac{v_{mn} v_{nk}}{W_n - W_k} \frac{1 - e^{2\pi i (W_m - W_n)t/\hbar}}{W_m - W_n}. \quad (21) \end{aligned}$$

We may suppose the diagonal elements  $v_{nn}$  of the perturbing energy to be zero, since if they were not zero they could be included with the proper energy  $W_n$ . There will then be no terms in (21) with vanishing denominators, provided all the energy levels are different.

Suppose now that the proper energy of the state  $m$  is equal to that of the initial state  $k$ . Then the first term on the right-hand side of (21) ceases to be periodic in the time, and becomes

$$\{v_{mk} - \sum_n v_{mn} v_{nk} / (W_n - W_k)\} 2\pi t / i\hbar,$$

which increases linearly with the time. The rate of increase consists of a part, proportional to  $v_{mk}$ , that is due to direct transitions from state  $k$ , together with a sum of parts, each of which is proportional to a  $v_{mn} v_{nk}$ , and is due to transitions first from  $k$  to  $n$  and then from  $n$  to  $m$ , although the amplitude  $a_n$  of the eigenfunction of the intermediate state always remains small.

When one applies the theory to the scattering of radiation one must consider not a single final state with exactly the same proper energy as the initial state, but a set of final states with proper energies lying close together in a range that contains the initial proper energy, corresponding to all the possible scattered light-quanta with different frequencies but the same direction of motion that

may appear. One must now determine the total probability of the system lying in any one of these final states, which is

$$\Sigma |a_m|^2 = \int (\Delta W_m)^{-1} |a_m|^2 dW_m,$$

where  $\Delta W_m$  is the interval between the energy levels. The second term in the expression (21) for  $a_m$  may be neglected since it always remains small (except in the case of resonance which will be considered later) and hence

$$\Sigma |a_m|^2 = \int \left| v_{mk} - \Sigma_n \frac{v_{mn} v_{nk}}{W_n - W_k} \right|^2 \frac{2[1 - \cos 2\pi(W_m - W_k)t/\hbar]}{\Delta W_m \cdot (W_m - W_k)^2} dW_m.$$

If one assumes that the integral converges, so that for large values of  $t$  practically the whole of it is contributed by values of  $W_m$  close to  $W_k$ , one obtains

$$\Sigma |a_m|^2 = \frac{4\pi^2 t}{\hbar \Delta W_m} \left| v_{mk} - \Sigma_n \frac{v_{mn} v_{nk}}{W_n - W_k} \right|^2, \quad (22)$$

where the quantities on the right refer to that final state that has exactly the initial proper energy.

We take the states  $k$  and  $m$  to be the same as for the true scattering process considered in the preceding section, so that equations (17), (18) and (19) still hold, and  $\Delta W_m = \hbar \Delta \nu_r = \hbar / \sigma_r \Delta \omega_r$ . We can now take the state  $n$  to be either the state  $J = J''$ ,  $N_s = N_s' - 1$ ,  $N_t = 0$  ( $t \neq s$ ) for any  $J''$ , which would make the process  $k \rightarrow n$  an absorption of an  $s$ -quantum and  $n \rightarrow m$  an emission of an  $r$ -quantum, or the state  $J = J''$ ,  $N_s = N_s'$ ,  $N_r = 1$ ,  $N_t = 0$  ( $t \neq s, r$ ), which would make  $k \rightarrow n$  the emission and  $n \rightarrow m$  the absorption. In the first case we should have

$$v_{nk} = \frac{e}{c} \left( \frac{\hbar \nu_s}{2\pi c \sigma_s} \right)^{\frac{1}{2}} \dot{x}_s (J'' J') N_s'^{\frac{1}{2}} \quad v_{mn} = \frac{e}{c} \left( \frac{\hbar \nu_r}{2\pi c \sigma_r} \right)^{\frac{1}{2}} \dot{x}_r (J' J''),$$

and

$$W_n = H_0(J'') + (N_s' - 1) \hbar \nu_s \quad W_n - W_k = \hbar [\nu(J'' J') - \nu_s]^*,$$

and in the second

$$v_{nk} = \frac{e}{c} \left( \frac{\hbar \nu_r}{2\pi c \sigma_r} \right)^{\frac{1}{2}} \dot{x}_r (J'' J') \quad v_{mn} = \frac{e}{c} \left( \frac{\hbar \nu_s}{2\pi c \sigma_s} \right)^{\frac{1}{2}} \dot{x}_s (J' J'') N_s'^{\frac{1}{2}},$$

and

$$W_n = H_0(J') + N_s' \hbar \nu_s + \hbar \nu_r \quad W_n - W_k = \hbar [\nu(J'' J') + \nu_r].$$

We shall neglect the other possible states  $n$ , namely those for which the matrix elements  $v_{mn}$ ,  $v_{nk}$  come from terms in the double summation in the Hamiltonian

\* The frequency  $\nu(J'' J')$  is not necessarily positive.

(13), as we are working only to the first order in these terms. (We are working to the second order only in the emission and absorption terms, which, as we shall find, is the same as the first order in the terms of the double summation.) We now obtain for the right-hand side of (22) in which we must take  $\nu_r = \nu_s$ ,

$$N'_s t \Delta \omega_r \frac{e^4 \nu_s^2}{h^2 c^6 \sigma_s} \left| \frac{\hbar}{m} \cos \alpha_{rs} - \sum_{J''} \left\{ \frac{\dot{x}_r(J'J'') \dot{x}_s(J''J')}{{\nu}(J''J') - \nu_s} + \frac{\dot{x}_s(J'J'') \dot{x}_r(J''J')}{{\nu}(J''J') + \nu_s} \right\} \right|^2 \quad (23)$$

The most convenient way of expressing this result is to find the amplitude  $P$  (a vector) of the electric moment of that vibrating dipole of frequency  $\nu_s$  that would, according to the classical theory, emit the same distribution of radiation as that actually scattered by the atom. The number of light-quanta of the type  $r$  (with  $\nu_r = \nu_s$ ) emitted by the dipole  $P$  in time  $t$  per unit solid angle is

$$2\pi^3 \nu_s^3 / hc^3 \cdot P_r^2 t,$$

where  $P_r$  is the component of  $P$  in the direction of the electric vector of the light-quanta  $r$ . Comparing this with (23) (which must first be divided by  $\Delta \omega_r$  to change it to the probability of a light quantum being scattered per unit solid angle) one finds for  $P_r$

$$\begin{aligned} P_r &= \left( \frac{8\pi N'_s}{hc^3 \nu_s \sigma_s} \right)^{\frac{1}{2}} \frac{e^2}{4\pi^2} \left| \frac{\hbar}{m} \cos \alpha_{rs} - \sum_{J''} \left\{ \frac{\dot{x}_r(J'J'') \dot{x}_s(J''J')}{\nu(J''J') - \nu_s} + \frac{\dot{x}_s(J'J'') \dot{x}_r(J''J')}{\nu(J''J') + \nu_s} \right\} \right| \\ &= E \frac{e^2}{h} \frac{1}{\nu_s^2} \left| \frac{\hbar}{4\pi^2 m} \cos \alpha_{rs} - \sum_{J''} [\nu(J''J')]^2 \left\{ \frac{x_r(J'J'') x_s(J''J')}{\nu(J''J') - \nu_s} \right. \right. \\ &\quad \left. \left. + \frac{x_s(J'J'') x_r(J''J')}{\nu(J''J') + \nu_s} \right\} \right|, \end{aligned} \quad (24)$$

using (7), where  $E$  is the amplitude of the electric vector of the incident radiation.

We can put this result in a different form by using the following relations, which follow from the quantum conditions,

$$\sum_{J''} [x_r(J'J'') x_s(J''J') - x_s(J'J'') x_r(J''J')] = [x_r x_s - x_s x_r](J'J') = 0 \quad (25)$$

and

$$\begin{aligned} \sum_{J''} [\dot{x}_r(J'J'') \dot{x}_s(J''J') - \dot{x}_s(J'J'') \dot{x}_r(J''J')] &= [x_r \dot{x}_s - \dot{x}_s x_r](J'J') \\ &= i\hbar / 2\pi m \cdot \cos \alpha_{rs}, \end{aligned} \quad (26)$$

which gives

$$\begin{aligned} \sum_{J''} [x_r(J'J'') x_s(J''J') \nu(J''J') + x_s(J'J'') x_r(J''J') \nu(J''J')] \\ &= \hbar / 4\pi^2 m \cdot \cos \alpha_{rs}. \end{aligned} \quad (27)$$

Multiplying (25) by  $\nu_s$  and adding to (27), we obtain

$$\begin{aligned} \Sigma_{J''} [x_r(J'J'')x_s(J''J')\{\nu(J''J') + \nu_s\} + x_s(J'J'')x_r(J''J')\{\nu(J''J') - \nu_s\}] \\ = h/4\pi^2 m \cdot \cos \alpha_{rs}. \end{aligned} \quad (28)$$

With the help of this equation, (24) reduces to

$$P_r = E \frac{e^2}{h} \left| \Sigma_{J''} \left\{ \frac{x_r(J'J'')x_s(J''J')}{\nu(J''J') - \nu_s} + \frac{x_s(J'J'')x_r(J''J')}{\nu(J''J') + \nu_s} \right\} \right|,$$

so that the vector  $P$  is equal to

$$P = E \frac{e^2}{h} \left| \Sigma_{J''} \left\{ \frac{x(J'J'')x_s(J''J')}{\nu(J''J') - \nu_s} + \frac{x_s(J'J'')x(J''J')}{\nu(J''J') + \nu_s} \right\} \right|, \quad (29)$$

where  $x$  without a suffix means the vector  $(x, y, z)$ . This is identical with Kramers' and Heisenberg's result.\*

In applying the formula (22), instead of taking the final state  $m$  of the system to be one for which the atom is again in its initial state  $J = J'$ , we can take a new final state for the atom,  $J = J'''$  say. The frequency  $\nu_r$  for the scattered radiation that gives no change of total proper energy is now

$$\nu_r = \nu_s - \nu(J'''J') = \nu_s + \nu(J''J'') - \nu(J''J'), \quad (30)$$

which differs from the incident frequency  $\nu_s$ , so that we obtain in this way the non-coherent scattered radiation. (We assume that this  $\nu_r$  is positive as otherwise there would be no non-coherent scattered radiation associated with the final state  $J = J'''$  of the atom.) In the present case we have  $v_{mk} = 0$ , corresponding to the fact that the true scattering process does not contribute to the non-coherent radiation. We now obtain for  $P_r$ , after a similar and almost identical calculation to that leading to equation (24),

$$\begin{aligned} P_r = E \frac{e^2}{h} \frac{1}{\nu_r \nu_s} \left| \Sigma_{J''} \nu(J''J') \nu(J''J'') \right. \\ \left. \left\{ \frac{x_r(J'''J'')x_s(J''J')}{\nu(J''J') - \nu_s} + \frac{x_s(J'''J'')x_r(J''J')}{\nu(J''J') + \nu_r} \right\} \right| \end{aligned} \quad (31)$$

This result can be put in the form corresponding to (29) with the help of equations analogous to (25) and (26) referring to the non-diagonal ( $J'''J'$ ) matrix elements of  $[x_r x_s - x_s x_r]$  and  $[x_r \dot{x}_s - \dot{x}_s x_r]$ . These equations give, corresponding to (28),

$$\Sigma_{J''} [x_r(J''J'')x_s(J''J')\{\nu(J''J') + \nu_s\} + x_s(J''J'')x_r(J''J')\{\nu(J''J'') - \nu_r\}] = 0.$$

\* Kramers and Heisenberg, *loc. cit.*, equation (18). For previous quantum-theoretical deductions of the dispersion formula see Born, Heisenberg and Jordan, 'Z. f. Physik,' vol. 35, p. 557, Kap. I, equation (40) (1926); Schrödinger, *loc. cit.*, § 2, equation (23); and Klein, *loc. cit.*, § 5, equation (82).

When the left-hand side of this equation is subtracted from the summation in (31) one obtains, on account of the relations

$$\begin{aligned}\nu(J''J')\nu(J''J''') &= \nu(J''J')[\nu(J''J') + \nu_r - \nu_s] \\ &= [\nu(J''J') - \nu_s][\nu(J''J') + \nu_r] + \nu_r\nu_s,\end{aligned}$$

and

$$\nu(J''J')\nu(J''J''') = [\nu(J''J''') - \nu_r][\nu(J''J') + \nu_r] + \nu_r\nu_s$$

which follow from (30), the result

$$P_r = E \frac{e^2}{\hbar} \left| \sum_{J''} \left\{ \frac{x_r(J'''J'')x_s(J''J')}{\nu(J''J') - \nu_s} + \frac{x_s(J'''J'')x_r(J''J')}{\nu(J''J') + \nu_r} \right\} \right|,$$

again in agreement with Kramers and Heisenberg.

### § 6. The Case of Resonance.

The dispersion formulæ obtained in the preceding section can no longer hold when the frequency of the incident radiation coincides with that of an absorption or emission line of the atom, on account of a vanishing denominator. One easily sees where a modification must be made in the deduction of the formulæ. Since one of the intermediate states  $n$  now has the same energy as the initial state  $k$ , the term in the second summation in (21) referring to this  $n$  becomes large and can no longer be neglected.

In investigating this case of resonance one must, for generality, suppose the incident radiation to consist of a distribution of light-quanta over a range of frequencies including the resonance frequency, instead of entirely of light-quanta of a single frequency, as the results will depend very considerably on how nearly monochromatic the incident radiation is. Thus one must take the initial state  $k$  of the system to be given by  $J = J'$  and  $N_s = N'_s$ , where  $N'_s$  is zero except for light-quanta of a specified direction, and is for these light-quanta (roughly speaking) a continuous function of the frequency, so that the rate of flow of incident energy per unit area per unit frequency range is given by (5). The final state  $m$  for a process of coherent scattering is one for which  $J = J'$  again, and a light-quantum  $s$  has been absorbed and one  $r$  of approximately the same frequency emitted. Thus we have

$$W_m - W_k = h(\nu_r - \nu_s). \quad (32)$$

As before, the intermediate states  $n$  will be those for which  $J = J''$  (arbitrary) and either the  $s$ -quantum has already been absorbed or the  $r$ -quantum has already been emitted. If we take for definiteness the case when the range of incident frequencies includes only one resonance frequency, and this is an

absorption frequency to the state of the atom  $J = J^l$ , say, then that intermediate state of the system for which  $J = J^l$  and for which the s-quantum has already been absorbed will have very nearly the same proper energy as the initial state. Calling this intermediate state  $l$  we have

$$W_l - W_k = \hbar (\nu_0 - \nu_s) \quad W_m - W_l = \hbar (\nu_r - \nu_0) \quad (33)$$

where  $\nu_0$  is the resonance frequency, equal to  $[H(J^l) - H(J')]/\hbar$ .

In equation (21) we can now neglect only those terms of the second summation for which  $n \neq l$ . This gives

$$\begin{aligned} a_m = & \left( v_{mk} - \sum_{n \neq l} \frac{v_{mn} v_{nk}}{W_n - W_k} \right) \frac{1 - e^{2\pi i (W_m - W_k) t/\hbar}}{W_m - W_k} \\ & + \frac{v_{ml} v_{lk}}{W_l - W_k} \left\{ \frac{1 - e^{2\pi i (W_m - W_l) t/\hbar}}{W_m - W_l} - \frac{1 - e^{2\pi i (W_m - W_k) t/\hbar}}{W_m - W_k} \right\}, \end{aligned}$$

which, with the help of (32) and (33), may be written

$$\begin{aligned} a_m = & \left( v_{mk} - \sum_{n \neq l} \frac{v_{mn} v_{nk}}{W_n - W_k} \right) \frac{1 - e^{2\pi i (\nu_r - \nu_s) t}}{\hbar (\nu_r - \nu_s)} \\ & + \frac{v_{ml} v_{lk}}{\hbar^2 (\nu_0 - \nu_s)} \left\{ \frac{1 - e^{2\pi i (\nu_r - \nu_0) t}}{\nu_r - \nu_0} - \frac{1 - e^{2\pi i (\nu_r - \nu_s) t}}{\nu_r - \nu_s} \right\}. \end{aligned}$$

We must now determine the total probability of a specified light-quantum  $r$  being emitted with the absorption of any one of the incident light-quanta  $s$ , which is given by  $\sum_{\nu_s} |a_m|^2$ , equal to  $\int (\Delta \nu_s)^{-1} |a_m|^2 d\nu_s$ . To evaluate this we require the following integrals

$$\begin{aligned} \int_0^\infty \frac{|1 - e^{2\pi i (\nu_r - \nu_s) t}|^2}{(\nu_r - \nu_s)^2} d\nu_s &= 4\pi^2 t \\ \int_0^\infty \frac{1}{(\nu_0 - \nu_s)^2} \left| \frac{1 - e^{2\pi i (\nu_r - \nu_0) t}}{\nu_r - \nu_0} - \frac{1 - e^{2\pi i (\nu_r - \nu_s) t}}{\nu_r - \nu_s} \right|^2 d\nu_s &= 4\pi \frac{2\pi (\nu_r - \nu_0) t - \sin 2\pi (\nu_r - \nu_0) t}{(\nu_r - \nu_0)^3} \\ \int_0^\infty \frac{1 - e^{2\pi i (\nu_r - \nu_s) t}}{(\nu_r - \nu_s) (\nu_0 - \nu_s)} \left\{ \frac{1 - e^{-2\pi i (\nu_r - \nu_0) t}}{\nu_r - \nu_0} - \frac{1 - e^{-2\pi i (\nu_r - \nu_s) t}}{\nu_r - \nu_s} \right\} d\nu_s &= 2\pi \left\{ \frac{2\pi (\nu_r - \nu_0) t - \sin 2\pi (\nu_r - \nu_0) t}{(\nu_r - \nu_0)^2} + i \frac{1 - \cos 2\pi (\nu_r - \nu_0) t}{(\nu_r - \nu_0)^2} \right\}, \end{aligned}$$

for large  $t$ , and with their help obtain,

$$\begin{aligned} \Sigma_{\nu_s} |a_m|^2 &= \left| v_{mk} - \Sigma_{n \neq l} \frac{v_{mn} v_{nk}}{W_n - W_k} \right|^2 \frac{4\pi^2 t}{h^2 \Delta \nu_s} \\ &+ \frac{|v_{ml} v_{lk}|^2}{h^4} \frac{4\pi}{\Delta \nu_s} \frac{2\pi (\nu_r - \nu_o) t - \sin 2\pi (\nu_r - \nu_o) t}{(\nu_r - \nu_o)^3} \\ &+ 2R \left( v_{mk} - \Sigma_{n \neq l} \frac{v_{mn} v_{nk}}{W_n - W_k} \right) \frac{v_{kl} v_{lm}}{h^3 \Delta \nu_s} 2\pi \left\{ \frac{2\pi (\nu_r - \nu_o) t - \sin 2\pi (\nu_r - \nu_o) t}{(\nu_r - \nu_o)^2} \right. \\ &\quad \left. + i \frac{1 - \cos 2\pi (\nu_r - \nu_o) t}{(\nu_r - \nu_o)^2} \right\} \quad (34) \end{aligned}$$

where the quantities on the right now refer to that incident light-quantum  $s$  for which  $\nu_s = \nu_r$ , and  $R$  means the real part of all that occurs in the term after it.

The first of these three terms is just the contribution of those terms of the dispersion formula (22) that remain finite, the second is that which replaces the contribution of the infinite term,\* and the third gives the interference between the first two, and replaces the cross terms obtained when one squares the dispersion electric moment. One can see the meaning of the second term more clearly if one sums it for all frequencies  $\nu_r$  of the scattered radiation in a small frequency range  $\nu_0 - \alpha'$  to  $\nu_0 + \alpha''$  about the resonance frequency  $\nu_0$  (which frequency range must be large compared with the theoretical breadth of the spectral line in order that the approximations may be valid). This is equivalent to multiplying the term by  $(\Delta \nu_r)^{-1}$  and integrating through the frequency range. If, for brevity, one denotes the quantity  $4\pi |v_{ml} v_{lk}|^2 / h^4 \Delta \nu_r \Delta \nu_s$  by  $f(\nu_r)$ , the result is, neglecting terms that do not increase indefinitely with  $t$  or that tend to zero as the  $\alpha$ 's tend to zero,

$$\begin{aligned} &\int_{\nu_0 - \alpha'}^{\nu_0 + \alpha''} f(\nu_r) \frac{2\pi (\nu_r - \nu_0) t - \sin 2\pi (\nu_r - \nu_0) t}{(\nu_r - \nu_0)^3} d\nu_r \\ &= f(\nu_0) \int_{\nu_0 - \alpha'}^{\nu_0 + \alpha''} \frac{2\pi (\nu_r - \nu_0) t - \sin 2\pi (\nu_r - \nu_0) t}{(\nu_r - \nu_0)^3} d\nu_r \\ &\quad + f'(\nu_0) \int_{\nu_0 - \alpha'}^{\nu_0 + \alpha''} \frac{2\pi (\nu_r - \nu_0) t - \sin 2\pi (\nu_r - \nu_0) t}{(\nu_r - \nu_0)^2} d\nu_r \\ &= f(\nu_0) (2\pi t)^2 \left[ \frac{1}{2} \pi - \frac{1}{2\pi t \alpha''} - \frac{1}{2\pi t \alpha'} \right] + f'(\nu_0) 2\pi t \log \alpha''/\alpha'. \end{aligned}$$

\* It should be noticed that this second term does not reduce to the square of the  $l$  term in the summation (22) when  $\nu_r$  is not a resonance frequency, but to double this amount. This difference is due to the fact that processes involving a change of proper energy are not entirely negligible for the initial conditions used in the present paper, and one such scattering process, which was neglected in § 5, becomes in the resonance case a process with no change of proper energy and is included in the calculation.

Thus the contribution of the second term in (34) to the small frequency range  $\nu_0 - \alpha'$  to  $\nu_0 + \alpha''$  consists of two parts, one of which increases proportionally to  $t^2$  and the other proportionally to  $t$ . The part that increases proportionally to  $t^2$ , namely,

$$\frac{1}{2}\pi f(\nu_0)(2\pi t)^2 = \frac{1}{2}(2\pi)^4 |v_m v_{lk}|^2 / h^4 \Delta\nu_r \Delta\nu_s \cdot t^2,$$

is just that which would arise from actual transitions to the higher state of the atom and down again governed by Einstein's laws, since the probability that the atom has been raised to the higher state by the time  $\tau$  is\*  $(2\pi)^2 |v_{lk}|^2 / h^2 \Delta\nu_s \cdot \tau$ , and when it is in the higher state the probability per unit time of its jumping down again with emission of a light-quantum in the required direction is  $(2\pi)^2 |v_{ml}|^2 / h^2 \Delta\nu_r$ , so that the total probability of the two transitions taking place within a time  $t$  is

$$\frac{(2\pi)^2 |v_{lk}|^2}{h^2 \Delta\nu_s} \cdot \frac{(2\pi)^2 |v_{ml}|^2}{h^2 \Delta\nu_r} \int_0^t \tau d\tau = \frac{(2\pi)^4 |v_m v_{lk}|^2}{h^4 \Delta\nu_r \Delta\nu_s} \frac{1}{2} t^2.$$

The part that increases linearly with the time may be added to the contributions of the first and third terms, which also increase according to this law. For values of  $t$  large compared with the periods of the atom, the terms proportional to  $t$  will be negligible compared with those proportional to  $t^2$ , and hence the resonance scattered radiation is due practically entirely to absorptions and emissions according to Einstein's laws.

---

\* This result and the one for the emission follow at once from formula (32) of *loc. cit.*



## *The Quantum Theory of the Electron.*

By P. A. M. DIRAC, St. John's College, Cambridge.

(Communicated by R. H. Fowler, F.R.S.—Received January 2, 1928.)

The new quantum mechanics, when applied to the problem of the structure of the atom with point-charge electrons, does not give results in agreement with experiment. The discrepancies consist of “duplexity” phenomena, the observed number of stationary states for an electron in an atom being twice the number given by the theory. To meet the difficulty, Goudsmit and Uhlenbeck have introduced the idea of an electron with a spin angular momentum of half a quantum and a magnetic moment of one Bohr magneton. This model for the electron has been fitted into the new mechanics by Pauli,\* and Darwin,† working with an equivalent theory, has shown that it gives results in agreement with experiment for hydrogen-like spectra to the first order of accuracy.

The question remains as to why Nature should have chosen this particular model for the electron instead of being satisfied with the point-charge. One would like to find some incompleteness in the previous methods of applying quantum mechanics to the point-charge electron such that, when removed, the whole of the duplexity phenomena follow without arbitrary assumptions. In the present paper it is shown that this is the case, the incompleteness of the previous theories lying in their disagreement with relativity, or, alternatively, with the general transformation theory of quantum mechanics. It appears that the simplest Hamiltonian for a point-charge electron satisfying the requirements of both relativity and the general transformation theory leads to an explanation of all duplexity phenomena without further assumption. All the same there is a great deal of truth in the spinning electron model, at least as a first approximation. The most important failure of the model seems to be that the magnitude of the resultant orbital angular momentum of an electron moving in an orbit in a central field of force is not a constant, as the model leads one to expect.

\* Pauli, ‘Z. f. Physik,’ vol. 43, p. 601 (1927).

† Darwin, ‘Roy. Soc. Proc.,’ A, vol. 116, p. 227 (1927).

§ 1. Previous Relativity Treatments.

The relativity Hamiltonian according to the classical theory for a point electron moving in an arbitrary electro-magnetic field with scalar potential  $A_0$  and vector potential  $\mathbf{A}$  is

$$F \equiv \left( \frac{W}{c} + \frac{e}{c} A_0 \right)^2 + \left( \mathbf{p} + \frac{e}{c} \mathbf{A} \right)^2 + m^2 c^2,$$

where  $\mathbf{p}$  is the momentum vector. It has been suggested by Gordon\* that the operator of the wave equation of the quantum theory should be obtained from this  $F$  by the same procedure as in non-relativity theory, namely, by putting

$$\begin{aligned} W &= i\hbar \frac{\partial}{\partial t}, \\ p_r &= -i\hbar \frac{\partial}{\partial x_r}, \quad r = 1, 2, 3, \end{aligned}$$

in it. This gives the wave equation

$$F\psi \equiv \left[ \left( i\hbar \frac{\partial}{c \partial t} + \frac{e}{c} A_0 \right)^2 + \sum_r \left( -i\hbar \frac{\partial}{\partial x_r} + \frac{e}{c} A_r \right)^2 + m^2 c^2 \right] \psi = 0, \quad (1)$$

the wave function  $\psi$  being a function of  $x_1, x_2, x_3, t$ . This gives rise to two difficulties.

The first is in connection with the physical interpretation of  $\psi$ . Gordon, and also independently Klein,† from considerations of the conservation theorems, make the assumption that if  $\psi_m, \psi_n$  are two solutions

$$\rho_{mn} = -\frac{e}{2mc^2} \left\{ i\hbar \left( \psi_m \frac{\partial \psi_n}{\partial t} - \bar{\psi}_n \frac{\partial \psi_m}{\partial t} \right) + 2eA_0 \psi_m \bar{\psi}_n \right\}$$

and

$$\mathbf{I}_{mn} = -\frac{e}{2m} \left\{ -i\hbar (\psi_m \operatorname{grad} \bar{\psi}_n - \bar{\psi}_n \operatorname{grad} \psi_m) + 2 \frac{e}{c} \mathbf{A}_0 \psi_m \bar{\psi}_n \right\}$$

are to be interpreted as the charge and current associated with the transition  $m \rightarrow n$ . This appears to be satisfactory so far as emission and absorption of radiation are concerned, but is not so general as the interpretation of the non-relativity quantum mechanics, which has been developed‡ sufficiently to enable one to answer the question : What is the probability of any dynamical variable

\* Gordon, 'Z. f. Physik,' vol. 40, p. 117 (1926).

† Klein, 'Z. f. Physik,' vol. 41, p. 407 (1927).

‡ Jordan, 'Z. f. Physik,' vol. 40, p. 809 (1927); Dirac, 'Roy. Soc. Proc.,' A, vol. 113, p. 621 (1927).

at any specified time having a value lying between any specified limits, when the system is represented by a given wave function  $\psi_n$ ? The Gordon-Klein interpretation can answer such questions if they refer to the position of the electron (by the use of  $\rho_{nn}$ ), but not if they refer to its momentum, or angular momentum or any other dynamical variable. We should expect the interpretation of the relativity theory to be just as general as that of the non-relativity theory.

The general interpretation of non-relativity quantum mechanics is based on the transformation theory, and is made possible by the wave equation being of the form

$$(H - W) \psi = 0, \quad (2)$$

i.e., being linear in  $W$  or  $\partial/\partial t$ , so that the wave function at any time determines the wave function at any later time. The wave equation of the relativity theory must also be linear in  $W$  if the general interpretation is to be possible.

The second difficulty in Gordon's interpretation arises from the fact that if one takes the conjugate imaginary of equation (1), one gets

$$\left[ \left( -\frac{W}{c} + \frac{e}{c} A_0 \right)^2 + \left( -\mathbf{p} + \frac{e}{c} \mathbf{A} \right)^2 + m^2 c^2 \right] \psi = 0,$$

which is the same as one would get if one put  $-e$  for  $e$ . The wave equation (1) thus refers equally well to an electron with charge  $e$  as to one with charge  $-e$ . If one considers for definiteness the limiting case of large quantum numbers one would find that some of the solutions of the wave equation are wave packets moving in the way a particle of charge  $-e$  would move on the classical theory, while others are wave packets moving in the way a particle of charge  $e$  would move classically. For this second class of solutions  $W$  has a negative value. One gets over the difficulty on the classical theory by arbitrarily excluding those solutions that have a negative  $W$ . One cannot do this on the quantum theory, since in general a perturbation will cause transitions from states with  $W$  positive to states with  $W$  negative. Such a transition would appear experimentally as the electron suddenly changing its charge from  $-e$  to  $e$ , a phenomenon which has not been observed. The true relativity wave equation should thus be such that its solutions split up into two non-combining sets, referring respectively to the charge  $-e$  and the charge  $e$ .

In the present paper we shall be concerned only with the removal of the first of these two difficulties. The resulting theory is therefore still only an approximation, but it appears to be good enough to account for all the duplexity phenomena without arbitrary assumptions.

§ 2. The Hamiltonian for No Field.

Our problem is to obtain a wave equation of the form (2) which shall be invariant under a Lorentz transformation and shall be equivalent to (1) in the limit of large quantum numbers. We shall consider first the case of no field, when equation (1) reduces to

$$(-p_0^2 + \mathbf{p}^2 + m^2c^2)\psi = 0 \quad (3)$$

if one puts

$$p_0 = \frac{W}{c} = i\hbar \frac{\partial}{c\partial t}.$$

The symmetry between  $p_0$  and  $p_1, p_2, p_3$  required by relativity shows that, since the Hamiltonian we want is linear in  $p_0$ , it must also be linear in  $p_1, p_2$  and  $p_3$ . Our wave equation is therefore of the form

$$(p_0 + \alpha_1 p_1 + \alpha_2 p_2 + \alpha_3 p_3 + \beta)\psi = 0, \quad (4)$$

where for the present all that is known about the dynamical variables or operators  $\alpha_1, \alpha_2, \alpha_3, \beta$  is that they are independent of  $p_0, p_1, p_2, p_3$ , i.e., that they commute with  $t, x_1, x_2, x_3$ . Since we are considering the case of a particle moving in empty space, so that all points in space are equivalent, we should expect the Hamiltonian not to involve  $t, x_1, x_2, x_3$ . This means that  $\alpha_1, \alpha_2, \alpha_3, \beta$  are independent of  $t, x_1, x_2, x_3$ , i.e., that they commute with  $p_0, p_1, p_2, p_3$ . We are therefore obliged to have other dynamical variables besides the co-ordinates and momenta of the electron, in order that  $\alpha_1, \alpha_2, \alpha_3, \beta$  may be functions of them. The wave function  $\psi$  must then involve more variables than merely  $x_1, x_2, x_3, t$ .

Equation (4) leads to

$$\begin{aligned} 0 &= (-p_0 + \alpha_1 p_1 + \alpha_2 p_2 + \alpha_3 p_3 + \beta)(p_0 + \alpha_1 p_1 + \alpha_2 p_2 + \alpha_3 p_3 + \beta)\psi \\ &= [-p_0^2 + \Sigma \alpha_1^2 p_1^2 + \Sigma (\alpha_1 \alpha_2 + \alpha_2 \alpha_1) p_1 p_2 + \beta^2 + \Sigma (\alpha_1 \beta + \beta \alpha_1) p_1] \psi, \end{aligned} \quad (5)$$

where the  $\Sigma$  refers to cyclic permutation of the suffixes 1, 2, 3. This agrees with (3) if

$$\left. \begin{aligned} \alpha_r^2 &= 1, & \alpha_r \alpha_s + \alpha_s \alpha_r &= 0 \quad (r \neq s) \\ \beta^2 &= m^2 c^2, & \alpha_r \beta + \beta \alpha_r &= 0 \end{aligned} \right\} \quad r, s = 1, 2, 3.$$

If we put  $\beta = \alpha_4 mc$ , these conditions become

$$\alpha_\mu^2 = 1 \quad \alpha_\mu \alpha_\nu + \alpha_\nu \alpha_\mu = 0 \quad (\mu \neq \nu) \quad \mu, \nu = 1, 2, 3, 4. \quad (6)$$

We can suppose the  $\alpha_\mu$ 's to be expressed as matrices in some matrix scheme, the matrix elements of  $\alpha_\mu$  being, say,  $\alpha_\mu(\zeta' \zeta')$ . The wave function  $\psi$  must

now be a function of  $\zeta$  as well as  $x_1, x_2, x_3, t$ . The result of  $\alpha_\mu$  multiplied into  $\psi$  will be a function  $(\alpha_\mu \psi)$  of  $x_1, x_2, x_3, t, \zeta$  defined by

$$(\alpha_\mu \psi)(x, t, \zeta) = \sum_{\zeta'} \alpha_\mu(\zeta \zeta') \psi(x, t, \zeta').$$

We must now find four matrices  $\alpha_\mu$  to satisfy the conditions (6). We make use of the matrices

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

which Pauli introduced\* to describe the three components of spin angular momentum. These matrices have just the properties

$$\sigma_r^2 = 1 \quad \sigma_r \sigma_s + \sigma_s \sigma_r = 0, \quad (r \neq s), \quad (7)$$

that we require for our  $\alpha$ 's. We cannot, however, just take the  $\sigma$ 's to be three of our  $\alpha$ 's, because then it would not be possible to find the fourth. We must extend the  $\sigma$ 's in a diagonal manner to bring in two more rows and columns, so that we can introduce three more matrices  $\rho_1, \rho_2, \rho_3$  of the same form as  $\sigma_1, \sigma_2, \sigma_3$ , but referring to different rows and columns, thus :—

$$\sigma_1 = \begin{Bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{Bmatrix} \quad \sigma_2 = \begin{Bmatrix} 0 & -i & 0 & 0 \\ i & 0 & 0 & 0 \\ 0 & 0 & 0 & -i \\ 0 & 0 & i & 0 \end{Bmatrix} \quad \sigma_3 = \begin{Bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{Bmatrix},$$

$$\rho_1 = \begin{Bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{Bmatrix} \quad \rho_2 = \begin{Bmatrix} 0 & 0 & -i & 0 \\ 0 & 0 & 0 & -i \\ i & 0 & 0 & 0 \\ 0 & i & 0 & 0 \end{Bmatrix} \quad \rho_3 = \begin{Bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{Bmatrix}.$$

The  $\rho$ 's are obtained from the  $\sigma$ 's by interchanging the second and third rows, and the second and third columns. We now have, in addition to equations (7)

$$\text{and also } \left. \begin{aligned} \rho_r^2 &= 1 & \rho_r \rho_s + \rho_s \rho_r &= 0 & (r \neq s), \\ \rho_r \sigma_t &= \sigma_t \rho_r. \end{aligned} \right\}. \quad (7')$$

\* Pauli, *loc. cit.*

If we now take

$$\alpha_1 = \rho_1\sigma_1, \quad \alpha_2 = \rho_1\sigma_2, \quad \alpha_3 = \rho_1\sigma_3, \quad \alpha_4 = \rho_3,$$

all the conditions (6) are satisfied, e.g.,

$$\begin{aligned} \alpha_1^2 &= \rho_1\sigma_1\rho_1\sigma_1 = \rho_1^2\sigma_1^2 = 1 \\ \alpha_1\alpha_2 &= \rho_1\sigma_1\rho_1\sigma_2 = \rho_1^2\sigma_1\sigma_2 = -\rho_1^2\sigma_2\sigma_1 = -\alpha_2\alpha_1. \end{aligned}$$

The following equations are to be noted for later reference

$$\left. \begin{aligned} \rho_1\rho_2 &= i\rho_3 = -\rho_2\rho_1 \\ \sigma_1\sigma_2 &= i\sigma_3 = -\sigma_2\sigma_1 \end{aligned} \right\}, \quad (8)$$

together with the equations obtained by cyclic permutation of the suffixes.

The wave equation (4) now takes the form

$$[p_0 + \rho_1(\sigma, p) + \rho_3mc] \psi = 0, \quad (9)$$

where  $\sigma$  denotes the vector  $(\sigma_1, \sigma_2, \sigma_3)$ .

### § 3. Proof of Invariance under a Lorentz Transformation.

Multiply equation (9) by  $\rho_3$  on the left-hand side. It becomes, with the help of (8),

$$[\rho_3p_0 + i\rho_2(\sigma_1p_1 + \sigma_2p_2 + \sigma_3p_3) + mc] \psi = 0.$$

Putting

$$p_0 = ip_4,$$

$$\rho_3 = \gamma_4, \quad \rho_2\sigma_r = \gamma_r, \quad r = 1, 2, 3, \quad (10)$$

we have

$$[i\sum\gamma_\mu p_\mu + mc] \psi = 0, \quad \mu = 1, 2, 3, 4. \quad (11)$$

The  $p_\mu$  transform under a Lorentz transformation according to the law

$$p_\mu' = \Sigma_\nu a_{\mu\nu} p_\nu,$$

where the coefficients  $a_{\mu\nu}$  are c-numbers satisfying

$$\Sigma_\mu a_{\mu\nu} a_{\mu\tau} = \delta_{\nu\tau}, \quad \Sigma_\tau a_{\mu\tau} a_{\nu\tau} = \delta_{\mu\nu}.$$

The wave equation therefore transforms into

$$[i\sum\gamma_\mu' p_\mu' + mc] \psi = 0, \quad (12)$$

where

$$\gamma_\mu' = \Sigma_\nu a_{\mu\nu} \gamma_\nu.$$

Now the  $\gamma_\mu$ , like the  $\alpha_\mu$ , satisfy

$$\gamma_\mu^2 = 1, \quad \gamma_\mu\gamma_\nu + \gamma_\nu\gamma_\mu = 0, \quad (\mu \neq \nu).$$

These relations can be summed up in the single equation

$$\gamma_\mu \gamma_\nu + \gamma_\nu \gamma_\mu = 2\delta_{\mu\nu}.$$

We have

$$\begin{aligned}\gamma_\mu' \gamma_\nu' + \gamma_\nu' \gamma_\mu' &= \sum_{\tau\lambda} a_{\mu\tau} a_{\nu\lambda} (\gamma_\tau \gamma_\lambda + \gamma_\lambda \gamma_\tau) \\ &= 2 \sum_{\tau\lambda} a_{\mu\tau} a_{\nu\lambda} \delta_{\tau\lambda} \\ &= 2 \sum_\tau a_{\mu\tau} a_{\nu\tau} = 2\delta_{\mu\nu}.\end{aligned}$$

Thus the  $\gamma'_\mu$ 's satisfy the same relations as the  $\gamma_\mu$ 's. Thus we can put, analogously to (10)

$$\gamma_4' = \rho_3' \quad \gamma_r' = \rho_2' \sigma_r'$$

where the  $\rho''$ 's and  $\sigma''$ 's are easily verified to satisfy the relations corresponding to (7), (7') and (8), if  $\rho_2'$  and  $\rho_1'$  are defined by  $\rho_2' = -i\gamma_1' \gamma_2' \gamma_3'$ ,  $\rho_1' = -i\rho_2' \rho_3'$ .

We shall now show that, by a canonical transformation, the  $\rho''$ 's and  $\sigma''$ 's may be brought into the form of the  $\rho$ 's and  $\sigma$ 's. From the equation  $\rho_3'^2 = 1$ , it follows that the only possible characteristic values for  $\rho_3'$  are  $\pm 1$ . If one applies to  $\rho_3'$  a canonical transformation with the transformation function  $\rho_1'$ , the result is

$$\rho_1' \rho_3' (\rho_1')^{-1} = -\rho_3' \rho_1' (\rho_1')^{-1} = -\rho_3'.$$

Since characteristic values are not changed by a canonical transformation,  $\rho_3'$  must have the same characteristic values as  $-\rho_3'$ . Hence the characteristic values of  $\rho_3'$  are  $+1$  twice and  $-1$  twice. The same argument applies to each of the other  $\rho''$ 's, and to each of the  $\sigma''$ 's.

Since  $\rho_3'$  and  $\sigma_3'$  commute, they can be brought simultaneously to the diagonal form by a canonical transformation. They will then have for their diagonal elements each  $+1$  twice and  $-1$  twice. Thus, by suitably rearranging the rows and columns, they can be brought into the form  $\rho_3$  and  $\sigma_3$  respectively. (The possibility  $\rho_3' = \pm \sigma_3'$  is excluded by the existence of matrices that commute with one but not with the other.)

Any matrix containing four rows and columns can be expressed as

$$c + \sum_r c_r \sigma_r + \sum_r c_r' \rho_r + \sum_{rs} c_{rs} \rho_r \sigma_s \quad (13)$$

where the sixteen coefficients  $c$ ,  $c_r$ ,  $c_r'$ ,  $c_{rs}$  are c-numbers. By expressing  $\sigma_1'$  in this way, we see, from the fact that it commutes with  $\rho_3' = \rho_3$  and anti-commutes\* with  $\sigma_3' = \sigma_3$ , that it must be of the form

$$\sigma_1' = c_1 \sigma_1 + c_2 \sigma_2 + c_{31} \rho_3 \sigma_1 + c_{32} \rho_3 \sigma_2,$$

\* We say that  $a$  anticommutes with  $b$  when  $ab = -ba$ .

i.e., of the form

$$\sigma_1' = \begin{Bmatrix} 0 & a_{12} & 0 & 0 \\ a_{21} & 0 & 0 & 0 \\ 0 & 0 & 0 & a_{34} \\ 0 & 0 & a_{43} & 0 \end{Bmatrix}.$$

The condition  $\sigma_1'^2 = 1$  shows that  $a_{12}a_{21} = 1$ ,  $a_{34}a_{43} = 1$ . If we now apply the canonical transformation : first row to be multiplied by  $(a_{21}/a_{12})^{\frac{1}{2}}$  and third row to be multiplied by  $(a_{43}/a_{34})^{\frac{1}{2}}$ , and first and third columns to be divided by the same expressions,  $\sigma_1'$  will be brought into the form of  $\sigma_1$ , and the diagonal matrices  $\sigma_3'$  and  $\rho_3'$  will not be changed.

If we now express  $\rho_1'$  in the form (13) and use the conditions that it commutes with  $\sigma_1' = \sigma_1$  and  $\sigma_3' = \sigma_3$  and anticommutes with  $\rho_3' = \rho_3$ , we see that it must be of the form

$$\rho_1' = c_1'\rho_1 + c_2'\rho_2.$$

The condition  $\rho_1'^2 = 1$  shows that  $c_1'^2 + c_2'^2 = 1$ , or  $c_1' = \cos \theta$ ,  $c_2' = \sin \theta$ . Hence  $\rho_1'$  is of the form

$$\rho_1' = \begin{Bmatrix} 0 & 0 & e^{-i\theta} & 0 \\ 0 & 0 & 0 & e^{-i\theta} \\ e^{i\theta} & 0 & 0 & 0 \\ 0 & e^{i\theta} & 0 & 0 \end{Bmatrix}$$

If we now apply the canonical transformation : first and second rows to be multiplied by  $e^{i\theta}$  and first and second columns to be divided by the same expression,  $\rho_1'$  will be brought into the form  $\rho_1$ , and  $\sigma_1$ ,  $\sigma_3$ ,  $\rho_3$  will not be altered.  $\rho_2'$  and  $\sigma_2'$  must now be of the form  $\rho_2$  and  $\sigma_2$ , on account of the relations  $i\rho_2' = \rho_3'\rho_1'$ ,  $i\sigma_2' = \sigma_3'\sigma_1'$ .

Thus by a succession of canonical transformations, which can be combined to form a single canonical transformation, the  $\rho$ 's and  $\sigma$ 's can be brought into the form of the  $\rho$ 's and  $\sigma$ 's. The new wave equation (12) can in this way be brought back into the form of the original wave equation (11) or (9), so that the results that follow from this original wave equation must be independent of the frame of reference used.

§ 4. *The Hamiltonian for an Arbitrary Field.*

To obtain the Hamiltonian for an electron in an electromagnetic field with scalar potential  $A_0$  and vector potential  $\mathbf{A}$ , we adopt the usual procedure of substituting  $p_0 + e/c \cdot A_0$  for  $p_0$  and  $\mathbf{p} + e/c \cdot \mathbf{A}$  for  $\mathbf{p}$  in the Hamiltonian for no field. From equation (9) we thus obtain

$$\left[ p_0 + \frac{e}{c} A_0 + \rho_1 \left( \boldsymbol{\sigma}, \mathbf{p} + \frac{e}{c} \mathbf{A} \right) + \rho_3 mc \right] \psi = 0. \quad (14)$$

This wave equation appears to be sufficient to account for all the duplexity phenomena. On account of the matrices  $\rho$  and  $\sigma$  containing four rows and columns, it will have four times as many solutions as the non-relativity wave equation, and twice as many as the previous relativity wave equation (1). Since half the solutions must be rejected as referring to the charge  $+e$  on the electron, the correct number will be left to account for duplexity phenomena. The proof given in the preceding section of invariance under a Lorentz transformation applies equally well to the more general wave equation (14).

We can obtain a rough idea of how (14) differs from the previous relativity wave equation (1) by multiplying it up analogously to (5). This gives, if we write  $e'$  for  $e/c$

$$\begin{aligned} 0 &= [-(p_0 + e'A_0) + \rho_1(\boldsymbol{\sigma}, \mathbf{p} + e'\mathbf{A}) + \rho_3 mc] \\ &\quad \times [(p_0 + e'A_0) + \rho_1(\boldsymbol{\sigma}, \mathbf{p} + e'\mathbf{A}) + \rho_3 mc] \psi \\ &= [-(p_0 + e'A_0)^2 + (\boldsymbol{\sigma}, \mathbf{p} + e'\mathbf{A})^2 + m^2 c^2 \\ &\quad + \rho_1\{(\boldsymbol{\sigma}, \mathbf{p} + e'\mathbf{A})(p_0 + e'A_0) - (p_0 + e'A_0)(\boldsymbol{\sigma}, \mathbf{p} + e'\mathbf{A})\}] \psi. \end{aligned} \quad (15)$$

We now use the general formula, that if  $\mathbf{B}$  and  $\mathbf{C}$  are any two vectors that commute with  $\boldsymbol{\sigma}$

$$\begin{aligned} (\boldsymbol{\sigma}, \mathbf{B})(\boldsymbol{\sigma}, \mathbf{C}) &= \Sigma \sigma_1^2 B_1 C_1 + \Sigma (\sigma_1 \sigma_2 B_1 C_2 + \sigma_2 \sigma_1 B_2 C_1) \\ &= (\mathbf{B}, \mathbf{C}) + i \Sigma \sigma_3 (B_1 C_2 - B_2 C_1) \\ &= (\mathbf{B}, \mathbf{C}) + i (\boldsymbol{\sigma}, \mathbf{B} \times \mathbf{C}). \end{aligned} \quad (16)$$

Taking  $\mathbf{B} = \mathbf{C} = \mathbf{p} + e'\mathbf{A}$ , we find

$$\begin{aligned} (\boldsymbol{\sigma}, \mathbf{p} + e'\mathbf{A})^2 &= (\mathbf{p} + e'\mathbf{A})^2 + i \Sigma \sigma_3 \\ &\quad [(p_1 + e'A_1)(p_2 + e'A_2) - (p_2 + e'A_2)(p_1 + e'A_1)] \\ &= (\mathbf{p} + e'\mathbf{A})^2 + h e' (\boldsymbol{\sigma}, \operatorname{curl} \mathbf{A}). \end{aligned}$$

Thus (15) becomes

$$\begin{aligned} 0 &= \left[ -(p_0 + e'A_0)^2 + (\mathbf{p} + e'\mathbf{A})^2 + m^2c^2 + e'h(\boldsymbol{\sigma}, \operatorname{curl} \mathbf{A}) \right. \\ &\quad \left. - ie'h\rho_1(\boldsymbol{\sigma}, \operatorname{grad} A_0 + \frac{1}{c} \frac{\partial \mathbf{A}}{\partial t}) \right] \psi \\ &= [-(p_0 + e'A_0)^2 + (\mathbf{p} + e'\mathbf{A})^2 + m^2c^2 + e'h(\boldsymbol{\sigma}, \mathbf{H}) + ie'h\rho_1(\boldsymbol{\sigma}, \mathbf{E})] \psi, \end{aligned}$$

where  $\mathbf{E}$  and  $\mathbf{H}$  are the electric and magnetic vectors of the field.

This differs from (1) by the two extra terms

$$\frac{eh}{c}(\boldsymbol{\sigma}, \mathbf{H}) + \frac{ieh}{c}\rho_1(\boldsymbol{\sigma}, \mathbf{E})$$

in  $F$ . These two terms, when divided by the factor  $2m$ , can be regarded as the additional potential energy of the electron due to its new degree of freedom. The electron will therefore behave as though it has a magnetic moment  $eh/2mc$ .  $\boldsymbol{\sigma}$  and an electric moment  $ieh/2mc$ .  $\rho_1\boldsymbol{\sigma}$ . This magnetic moment is just that assumed in the spinning electron model. The electric moment, being a pure imaginary, we should not expect to appear in the model. It is doubtful whether the electric moment has any physical meaning, since the Hamiltonian in (14) that we started from is real, and the imaginary part only appeared when we multiplied it up in an artificial way in order to make it resemble the Hamiltonian of previous theories.

### § 5. The Angular Momentum Integrals for Motion in a Central Field.

We shall consider in greater detail the motion of an electron in a central field of force. We put  $\mathbf{A} = 0$  and  $e'A_0 = V(r)$ , an arbitrary function of the radius  $r$ , so that the Hamiltonian in (14) becomes

$$F \equiv p_0 + V + \rho_1(\boldsymbol{\sigma}, \mathbf{p}) + \rho_3 mc.$$

We shall determine the periodic solutions of the wave equation  $F\psi = 0$ , which means that  $p_0$  is to be counted as a parameter instead of an operator; it is, in fact, just  $1/c$  times the energy level.

We shall first find the angular momentum integrals of the motion. The orbital angular momentum  $\mathbf{m}$  is defined by

$$\mathbf{m} = \mathbf{x} \times \mathbf{p},$$

and satisfies the following "Vertauschungs" relations

$$\left. \begin{aligned} m_1x_1 - x_1m_1 &= 0, & m_1x_2 - x_2m_1 &= ihx_3 \\ m_1p_1 - p_1m_1 &= 0, & m_1p_2 - p_2m_1 &= ihp_3 \\ \mathbf{m} \times \mathbf{m} &= ihm, & \mathbf{m}^2m_1 - m_1\mathbf{m}^2 &= 0, \end{aligned} \right\}, \quad (17)$$

together with similar relations obtained by permuting the suffixes. Also  $\mathbf{m}$  commutes with  $r$ , and with  $p_r$ , the momentum canonically conjugate to  $r$ .

We have

$$\begin{aligned} m_1 F - F m_1 &= \rho_1 \{ m_1 (\boldsymbol{\sigma}, \mathbf{p}) - (\boldsymbol{\sigma}, \mathbf{p}) m_1 \} \\ &= \rho_1 (\boldsymbol{\sigma}, m_1 \mathbf{p} - \mathbf{p} m_1) \\ &= i\hbar \rho_1 (\sigma_2 p_3 - \sigma_3 p_2), \end{aligned}$$

and so

$$\mathbf{m} F - F \mathbf{m} = i\hbar \rho_1 \boldsymbol{\sigma} \times \mathbf{p}. \quad (18)$$

Thus  $\mathbf{m}$  is not a constant of the motion. We have further

$$\begin{aligned} \sigma_1 F - F \sigma_1 &= \rho_1 \{ \sigma_1 (\boldsymbol{\sigma}, \mathbf{p}) - (\boldsymbol{\sigma}, \mathbf{p}) \sigma_1 \} \\ &= \rho_1 (\sigma_1 \boldsymbol{\sigma} - \boldsymbol{\sigma} \sigma_1, \mathbf{p}) \\ &= 2i\rho_1 (\sigma_3 p_2 - \sigma_2 p_3), \end{aligned}$$

with the help of (8), and so

$$\boldsymbol{\sigma} F - F \boldsymbol{\sigma} = -2i\rho_1 \boldsymbol{\sigma} \times \mathbf{p}.$$

Hence

$$(\mathbf{m} + \frac{1}{2}\hbar \boldsymbol{\sigma}) F - F (\mathbf{m} + \frac{1}{2}\hbar \boldsymbol{\sigma}) = 0.$$

Thus  $\mathbf{m} + \frac{1}{2}\hbar \boldsymbol{\sigma}$  ( $= \mathbf{M}$  say) is a constant of the motion. We can interpret this result by saying that the electron has a spin angular momentum of  $\frac{1}{2}\hbar \boldsymbol{\sigma}$ , which, added to the orbital angular momentum  $\mathbf{m}$ , gives the total angular momentum  $\mathbf{M}$ , which is a constant of the motion.

The Vertauschungs relations (17) all hold when  $M$ 's are written for the  $m$ 's.

In particular

$$\mathbf{M} \times \mathbf{M} = i\hbar \mathbf{M} \quad \text{and} \quad \mathbf{M}^2 \mathbf{M}_3 = \mathbf{M}_3 \mathbf{M}^2.$$

$\mathbf{M}_3$  will be an action variable of the system. Since the characteristic values of  $m_3$  must be integral multiples of  $\hbar$  in order that the wave function may be single-valued, the characteristic values of  $\mathbf{M}_3$  must be half odd integral multiples of  $\hbar$ . If we put

$$\mathbf{M}^2 = (j^2 - \frac{1}{4}) \hbar^2, \quad (19)$$

$j$  will be another quantum number, and the characteristic values of  $\mathbf{M}_3$  will extend from  $(j - \frac{1}{2}) \hbar$  to  $(-j + \frac{1}{2}) \hbar$ .\* Thus  $j$  takes integral values.

One easily verifies from (18) that  $\mathbf{m}^2$  does not commute with  $F$ , and is thus not a constant of the motion. This makes a difference between the present theory and the previous spinning electron theory, in which  $\mathbf{m}^2$  is constant, and defines the azimuthal quantum number  $k$  by a relation similar to (19). We shall find that our  $j$  plays the same part as the  $k$  of the previous theory.

\* See 'Roy. Soc. Proc.,' A, vol. 111, p. 281 (1926).

§ 6. The Energy Levels for Motion in a Central Field.

We shall now obtain the wave equation as a differential equation in  $r$ , with the variables that specify the orientation of the whole system removed. We can do this by the use only of elementary non-commutative algebra in the following way.

In formula (16) take  $\mathbf{B} = \mathbf{C} = \mathbf{m}$ . This gives

$$\begin{aligned} (\sigma, \mathbf{m})^2 &= \mathbf{m}^2 + i(\sigma, \mathbf{m} \times \mathbf{m}) \\ &= (\mathbf{m} + \frac{1}{2}\hbar\sigma)^2 - \hbar(\sigma, \mathbf{m}) - \frac{1}{4}\hbar^2\sigma^2 - \hbar(\sigma, \mathbf{m}) \\ &= \mathbf{M}^2 - 2\hbar(\sigma, \mathbf{m}) - \frac{3}{4}\hbar^2. \end{aligned} \quad (20)$$

Hence

$$\{(\sigma, \mathbf{m}) + \hbar\}^2 = \mathbf{M}^2 + \frac{1}{4}\hbar^2 = j^2\hbar^2.$$

Up to the present we have defined  $j$  only through  $j^2$ , so that we could now, if we liked, take  $j\hbar$  equal to  $(\sigma, \mathbf{m}) + \hbar$ . This would not be convenient since we want  $j$  to be a constant of the motion while  $(\sigma, \mathbf{m}) + \hbar$  is not, although its square is. We have, in fact, by another application of (16),

$$(\sigma, \mathbf{m})(\sigma, \mathbf{p}) = i(\sigma, \mathbf{m} \times \mathbf{p})$$

since  $(\mathbf{m}, \mathbf{p}) = 0$ , and similarly

$$(\sigma, \mathbf{p})(\sigma, \mathbf{m}) = i(\sigma, \mathbf{p} \times \mathbf{m}),$$

so that

$$\begin{aligned} (\sigma, \mathbf{m})(\sigma, \mathbf{p}) + (\sigma, \mathbf{p})(\sigma, \mathbf{m}) &= i\Sigma_{\sigma_1} (m_2 p_3 - m_3 p_2 + p_2 m_3 - p_3 m_2) \\ &= i\Sigma_{\sigma_1} \cdot 2i\hbar p_1 = -2\hbar(\sigma, \mathbf{p}), \end{aligned}$$

or

$$\{(\sigma, \mathbf{m}) + \hbar\}(\sigma, \mathbf{p}) + (\sigma, \mathbf{p})\{(\sigma, \mathbf{m}) + \hbar\} = 0.$$

Thus  $(\sigma, \mathbf{m}) + \hbar$  anticommutes with one of the terms in  $F$ , namely,  $\rho_1(\sigma, \mathbf{p})$ , and commutes with the other three. Hence  $\rho_3\{(\sigma, \mathbf{m}) + \hbar\}$  commutes with all four, and is therefore a constant of the motion. But the square of  $\rho_3\{(\sigma, \mathbf{m}) + \hbar\}$  must also equal  $j^2\hbar^2$ . We therefore take

$$j\hbar = \rho_3\{(\sigma, \mathbf{m}) + \hbar\}. \quad (21)$$

We have, by a further application of (16)

$$(\sigma, \mathbf{x})(\sigma, \mathbf{p}) = (\mathbf{x}, \mathbf{p}) + i(\sigma, \mathbf{m}).$$

Now a permissible definition of  $p_r$  is

$$(\mathbf{x}, \mathbf{p}) = r p_r + i\hbar,$$

and from (21)

$$(\sigma, \mathbf{m}) = \rho_3 j\hbar - \hbar.$$

Hence

$$(\sigma, \mathbf{x})(\sigma, \mathbf{p}) = r p_r + i\rho_3 j\hbar. \quad (22)$$

Introduce the quantity  $\varepsilon$  defined by

$$r\varepsilon = \rho_1(\sigma, \mathbf{x}). \quad (23)$$

Since  $r$  commutes with  $\rho_1$  and with  $(\sigma, \mathbf{x})$ , it must commute with  $\varepsilon$ . We thus have

$$r^2\varepsilon^2 = [\rho_1(\sigma, \mathbf{x})]^2 = (\sigma, \mathbf{x})^2 = \mathbf{x}^2 = r^2$$

or

$$\varepsilon^2 = 1.$$

Since there is symmetry between  $\mathbf{x}$  and  $\mathbf{p}$  so far as angular momentum is concerned,  $\rho_1(\sigma, \mathbf{x})$ , like  $\rho_1(\sigma, \mathbf{p})$ , must commute with  $\mathbf{M}$  and  $j$ . Hence  $\varepsilon$  commutes with  $\mathbf{M}$  and  $j$ . Further,  $\varepsilon$  must commute with  $p_r$ , since we have

$$(\sigma, \mathbf{x})(\mathbf{x}, \mathbf{p}) - (\mathbf{x}, \mathbf{p})(\sigma, \mathbf{x}) = ih(\sigma, \mathbf{x}),$$

which gives

$$r\varepsilon(r p_r + ih) - (r p_r + ih)r\varepsilon = i h r \varepsilon,$$

which reduces to

$$\varepsilon p_r - p_r \varepsilon = 0.$$

From (22) and (23) we now have

$$r\varepsilon\rho_1(\sigma, \mathbf{p}) = r p_r + i \rho_3 j h$$

or

$$\rho_1(\sigma, \mathbf{p}) = \varepsilon p_r + i \varepsilon \rho_3 j h / r.$$

Thus

$$F = p_0 + V + \varepsilon p_r + i \varepsilon \rho_3 j h / r + \rho_3 m c. \quad (24)$$

Equation (23) shows that  $\varepsilon$  anticommutes with  $\rho_3$ . We can therefore by a canonical transformation (involving perhaps the  $x$ 's and  $p$ 's as well as the  $\sigma$ 's and  $\rho$ 's) bring  $\varepsilon$  into the form of the  $\rho_2$  of § 2 without changing  $\rho_3$ , and without changing any of the other variables occurring on the right-hand side of (24), since these other variables all commute with  $\varepsilon$ .  $i \varepsilon \rho_3$  will now be of the form  $i \rho_2 \rho_3 = -\rho_1$ , so that the wave equation takes the form

$$F\psi \equiv [p_0 + V + \rho_2 p_r - \rho_1 j h / r + \rho_3 m c] \psi = 0.$$

If we write this equation out in full, calling the components of  $\psi$  referring to the first and third rows (or columns) of the matrices  $\psi_\alpha$  and  $\psi_\beta$  respectively, we get

$$(F\psi)_\alpha \equiv (p_0 + V) \psi_\alpha - h \frac{\partial}{\partial r} \psi_\beta - \frac{j h}{r} \psi_\beta + m c \psi_\alpha = 0,$$

$$(F\psi)_\beta \equiv (p_0 + V) \psi_\beta + h \frac{\partial}{\partial r} \psi_\alpha - \frac{j h}{r} \psi_\alpha - m c \psi_\beta = 0.$$

The second and fourth components give just a repetition of these two equations. We shall now eliminate  $\psi_a$ . If we write  $hB$  for  $p_0 + V + mc$ , the first equation becomes

$$\left( \frac{\partial}{\partial r} + \frac{j}{r} \right) \psi_\beta = B \psi_a,$$

which gives on differentiating

$$\begin{aligned} \frac{\partial^2}{\partial r^2} \psi_\beta + \frac{j}{r} \frac{\partial}{\partial r} \psi_\beta - \frac{j}{r^2} \psi_\beta &= B \frac{\partial}{\partial r} \psi_a + \frac{\partial B}{\partial r} \psi_a \\ &= \frac{B}{\hbar} \left[ -(p_0 + V - mc) \psi_\beta + \frac{j\hbar}{r} \psi_a \right] + \frac{1}{\hbar} \frac{\partial V}{\partial r} \psi_a \\ &= -\frac{(p_0 + V)^2 - m^2 c^2}{\hbar^2} \psi_\beta + \left( \frac{j}{r} + \frac{1}{B\hbar} \frac{\partial V}{\partial r} \right) \left( \frac{\partial}{\partial r} + \frac{j}{r} \right) \psi_\beta. \end{aligned}$$

This reduces to

$$\frac{\partial^2}{\partial r^2} \psi_\beta + \left[ \frac{(p_0 + V)^2 - m^2 c^2}{\hbar^2} - \frac{j(j+1)}{r^2} \right] \psi_\beta - \frac{1}{B\hbar} \frac{\partial V}{\partial r} \left( \frac{\partial}{\partial r} + \frac{j}{r} \right) \psi_\beta = 0. \quad (25)$$

The values of the parameter  $p_0$  for which this equation has a solution finite at  $r = 0$  and  $r = \infty$  are  $1/c$  times the energy levels of the system. To compare this equation with those of previous theories, we put  $\psi_\beta = r\chi$ , so that

$$\frac{\partial^2}{\partial r^2} \chi + \frac{2}{r} \frac{\partial}{\partial r} \chi + \left[ \frac{(p_0 + V)^2 - m^2 c^2}{\hbar^2} - \frac{j(j+1)}{r^2} \right] \chi - \frac{1}{B\hbar} \frac{\partial V}{\partial r} \left( \frac{\partial}{\partial r} + \frac{j+1}{r} \right) \chi = 0. \quad (26)$$

If one neglects the last term, which is small on account of  $B$  being large, this equation becomes the same as the ordinary Schroedinger equation for the system, with relativity correction included. Since  $j$  has, from its definition, both positive and negative integral characteristic values, our equation will give twice as many energy levels when the last term is not neglected.

We shall now compare the last term of (26), which is of the same order of magnitude as the relativity correction, with the spin correction given by Darwin and Pauli. To do this we must eliminate the  $\partial\chi/\partial r$  term by a further transformation of the wave function. We put

$$\chi = B^{-\frac{1}{2}} \chi_1,$$

which gives

$$\begin{aligned} \frac{\partial^2}{\partial r^2} \chi_1 + \frac{2}{r} \frac{\partial}{\partial r} \chi_1 + \left[ \frac{(p_0 + V)^2 - m^2 c^2}{\hbar^2} - \frac{j(j+1)}{r^2} \right] \chi_1 \\ + \left[ \frac{1}{B\hbar} \frac{j}{r} \frac{\partial V}{\partial r} - \frac{1}{2} \frac{1}{B\hbar} \frac{\partial^2 V}{\partial r^2} + \frac{1}{4} \frac{1}{B^2 \hbar^2} \left( \frac{\partial V}{\partial r} \right)^2 \right] \chi_1 = 0. \quad (27) \end{aligned}$$

The correction is now, to the first order of accuracy

$$\frac{1}{Bh} \left( \frac{j}{r} \frac{\partial V}{\partial r} - \frac{1}{2} \frac{\partial^2 V}{\partial r^2} \right),$$

where  $Bh = 2mc$  (provided  $p_0$  is positive). For the hydrogen atom we must put  $V = e^2/cr$ . The first order correction now becomes

$$-\frac{e^2}{2mc^2r^3}(j+1). \quad (28)$$

If we write  $-j$  for  $j+1$  in (27), we do not alter the terms representing the unperturbed system, so

$$\frac{e^2}{2mc^2r^3} j \quad (28')$$

will give a second possible correction for the same unperturbed term.

In the theory of Pauli and Darwin, the corresponding correcting term is

$$\frac{e^2}{2mhc^2r^3} (\sigma, m)$$

when the Thomas factor  $\frac{1}{2}$  is included. We must remember that in the Pauli-Darwin theory, the resultant orbital angular momentum  $k$  plays the part of our  $j$ . We must define  $k$  by

$$m^2 = k(k+1)h^2$$

instead of by the exact analogue of (19), in order that it may have integral characteristic values, like  $j$ . We have from (20)

$$(\sigma, m)^2 = k(k+1)h^2 - h(\sigma, m)$$

or

$$\{( \sigma, m ) + \frac{1}{2}h\}^2 = (k + \frac{1}{2})^2h^2,$$

hence

$$(\sigma, m) = kh \text{ or } -(k+1)h.$$

The correction thus becomes

$$\frac{e^2}{2mc^2r^3} k \quad \text{or} \quad -\frac{e^2}{2mc^2r^3}(k+1),$$

which agrees with (28) and (28'). The present theory will thus, in the first approximation, lead to the same energy levels as those obtained by Darwin, which are in agreement with experiment.



*A RELATION BETWEEN DISTANCE AND RADIAL VELOCITY  
AMONG EXTRA-GALACTIC NEBULAE*

BY EDWIN HUBBLE

MOUNT WILSON OBSERVATORY, CARNEGIE INSTITUTION OF WASHINGTON

Communicated January 17, 1929

Determinations of the motion of the sun with respect to the extra-galactic nebulae have involved a  $K$  term of several hundred kilometers which appears to be variable. Explanations of this paradox have been sought in a correlation between apparent radial velocities and distances, but so far the results have not been convincing. The present paper is a re-examination of the question, based on only those nebular distances which are believed to be fairly reliable.

Distances of extra-galactic nebulae depend ultimately upon the application of absolute-luminosity criteria to involved stars whose types can be recognized. These include, among others, Cepheid variables, novae, and blue stars involved in emission nebulosity. Numerical values depend upon the zero point of the period-luminosity relation among Cepheids, the other criteria merely check the order of the distances. This method is restricted to the few nebulae which are well resolved by existing instruments. A study of these nebulae, together with those in which any stars at all can be recognized, indicates the probability of an approximately uniform upper limit to the absolute luminosity of stars, in the late-type spirals and irregular nebulae at least, of the order of  $M$  (photographic) =  $-6.3$ .<sup>1</sup> The apparent luminosities of the brightest stars in such nebulae are thus criteria which, although rough and to be applied with caution,

furnish reasonable estimates of the distances of all extra-galactic systems in which even a few stars can be detected.

TABLE 1  
NEBULAE WHOSE DISTANCES HAVE BEEN ESTIMATED FROM STARS INVOLVED OR FROM  
MEAN LUMINOSITIES IN A CLUSTER

OBJECT	$m_s$	$r$	$v$	$m_t$	$M_t$
S. Mag.	..	0.032	+ 170	1.5	-16.0
L. Mag.	..	0.034	+ 290	0.5	17.2
N. G. C. 6822	..	0.214	- 130	9.0	12.7
598	..	0.263	- 70	7.0	15.1
221	..	0.275	- 185	8.8	13.4
224	..	0.275	- 220	5.0	17.2
5457	17.0	0.45	+ 200	9.9	13.3
4736	17.3	0.5	+ 290	8.4	15.1
5194	17.3	0.5	+ 270	7.4	16.1
4449	17.8	0.63	+ 200	9.5	14.5
4214	18.3	0.8	+ 300	11.3	13.2
3031	18.5	0.9	- 30	8.3	16.4
3627	18.5	0.9	+ 650	9.1	15.7
4826	18.5	0.9	+ 150	9.0	15.7
5236	18.5	0.9	+ 500	10.4	14.4
1068	18.7	1.0	+ 920	9.1	15.9
5055	19.0	1.1	+ 450	9.6	15.6
7331	19.0	1.1	+ 500	10.4	14.8
4258	19.5	1.4	+ 500	8.7	17.0
4151	20.0	1.7	+ 960	12.0	14.2
4382	..	2.0	+ 500	10.0	16.5
4472	..	2.0	+ 850	8.8	17.7
4486	..	2.0	+ 800	9.7	16.8
4649	..	2.0	+1090	9.5	17.0
Mean					-15.5

$m_s$  = photographic magnitude of brightest stars involved.

$r$  = distance in units of  $10^6$  parsecs. The first two are Shapley's values.

$v$  = measured velocities in km./sec. N. G. C. 6822, 221, 224 and 5457 are recent determinations by Humason.

$m_t$  = Holetschek's visual magnitude as corrected by Hopmann. The first three objects were not measured by Holetschek, and the values of  $m_t$  represent estimates by the author based upon such data as are available.

$M_t$  = total visual absolute magnitude computed from  $m_t$  and  $r$ .

Finally, the nebulae themselves appear to be of a definite order of absolute luminosity, exhibiting a range of four or five magnitudes about an average value  $M$  (visual) = -15.2.<sup>1</sup> The application of this statistical average to individual cases can rarely be used to advantage, but where considerable numbers are involved, and especially in the various clusters of nebulae, mean apparent luminosities of the nebulae themselves offer reliable estimates of the mean distances.

Radial velocities of 46 extra-galactic nebulae are now available, but

individual distances are estimated for only 24. For one other, N. G. C. 3521, an estimate could probably be made, but no photographs are available at Mount Wilson. The data are given in table 1. The first seven distances are the most reliable, depending, except for M 32 the companion of M 31, upon extensive investigations of many stars involved. The next thirteen distances, depending upon the criterion of a uniform upper limit of stellar luminosity, are subject to considerable probable errors but are believed to be the most reasonable values at present available. The last four objects appear to be in the Virgo Cluster. The distance assigned to the cluster,  $2 \times 10^6$  parsecs, is derived from the distribution of nebular luminosities, together with luminosities of stars in some of the later-type spirals, and differs somewhat from the Harvard estimate of ten million light years.<sup>2</sup>

The data in the table indicate a linear correlation between distances and velocities, whether the latter are used directly or corrected for solar motion, according to the older solutions. This suggests a new solution for the solar motion in which the distances are introduced as coefficients of the  $K$  term, i. e., the velocities are assumed to vary directly with the distances, and hence  $K$  represents the velocity at unit distance due to this effect. The equations of condition then take the form

$$rK + X \cos \alpha \cos \delta + Y \sin \alpha \cos \delta + Z \sin \delta = v.$$

Two solutions have been made, one using the 24 nebulae individually, the other combining them into 9 groups according to proximity in direction and in distance. The results are

	24 OBJECTS	9 GROUPS
$X$	$-65 \pm 50$	$+3 \pm 70$
$Y$	$+226 \pm 95$	$+230 \pm 120$
$Z$	$-195 \pm 40$	$-133 \pm 70$
$K$	$+465 \pm 50$	$+513 \pm 60 \text{ km./sec. per } 10^6 \text{ parsecs.}$
$A$	$286^\circ$	$269^\circ$
$D$	$+40^\circ$	$+33^\circ$
$V_0$	$306 \text{ km./sec.}$	$247 \text{ km./sec.}$

For such scanty material, so poorly distributed, the results are fairly definite. Differences between the two solutions are due largely to the four Virgo nebulae, which, being the most distant objects and all sharing the peculiar motion of the cluster, unduly influence the value of  $K$  and hence of  $V_0$ . New data on more distant objects will be required to reduce the effect of such peculiar motion. Meanwhile round numbers, intermediate between the two solutions, will represent the probable order of the values. For instance, let  $A = 277^\circ$ ,  $D = +36^\circ$  (Gal. long. =  $32^\circ$ , lat. =  $+18^\circ$ ),  $V_0 = 280 \text{ km./sec.}$ ,  $K = +500 \text{ km./sec. per million par-}$

secs. Mr. Strömgren has very kindly checked the general order of these values by independent solutions for different groupings of the data.

A constant term, introduced into the equations, was found to be small and negative. This seems to dispose of the necessity for the old constant  $K$  term. Solutions of this sort have been published by Lundmark,<sup>3</sup> who replaced the old  $K$  by  $k + lr + mr^2$ . His favored solution gave  $k = 513$ , as against the former value of the order of 700, and hence offered little advantage.

TABLE 2  
NEBULAE WHOSE DISTANCES ARE ESTIMATED FROM RADIAL VELOCITIES

OBJECT	$v$	$v_s$	$r$	$m_t$	$M_t$
N. G. C. 278	+ 650	-110	1.52	12.0	-13.9
404	- 25	- 65	..	11.1	..
584	+1800	+ 75	3.45	10.9	16.8
936	+1300	+115	2.37	11.1	15.7
1023	+ 300	- 10	0.62	10.2	13.8
1700	+ 800	+220	1.16	12.5	12.8
2681	+ 700	- 10	1.42	10.7	15.0
2683	+ 400	+ 65	0.67	9.9	14.3
2841	+ 600	- 20	1.24	9.4	16.1
3034	+ 290	-105	0.79	9.0	15.5
3115	+ 600	+105	1.00	9.5	15.5
3368	+ 940	+ 70	1.74	10.0	16.2
3379	+ 810	+ 65	1.49	9.4	16.4
3489	+ 600	+ 50	1.10	11.2	14.0
3521	+ 730	+ 95	1.27	10.1	15.4
3623	+ 800	+ 35	1.53	9.9	16.0
4111	+ 800	- 95	1.79	10.1	16.1
4526	+ 580	- 20	1.20	11.1	14.3
4565	+1100	- 75	2.35	11.0	15.9
4594	+1140	+ 25	2.23	9.1	17.6
5005	+ 900	-130	2.06	11.1	15.5
5866	+ 650	-215	1.73	11.7	-14.5
Mean				10.5	-15.3

The residuals for the two solutions given above average 150 and 110 km./sec. and should represent the average peculiar motions of the individual nebulae and of the groups, respectively. In order to exhibit the results in a graphical form, the solar motion has been eliminated from the observed velocities and the remainders, the distance terms plus the residuals, have been plotted against the distances. The run of the residuals is about as smooth as can be expected, and in general the form of the solutions appears to be adequate.

The 22 nebulae for which distances are not available can be treated in two ways. First, the mean distance of the group derived from the mean apparent magnitudes can be compared with the mean of the velocities

corrected for solar motion. The result, 745 km./sec. for a distance of  $1.4 \times 10^6$  parsecs, falls between the two previous solutions and indicates a value for  $K$  of 530 as against the proposed value, 500 km./sec.

Secondly, the scatter of the individual nebulae can be examined by assuming the relation between distances and velocities as previously determined. Distances can then be calculated from the velocities corrected for solar motion, and absolute magnitudes can be derived from the apparent magnitudes. The results are given in table 2 and may be compared with the distribution of absolute magnitudes among the nebulae in table 1, whose distances are derived from other criteria. N. G. C. 404

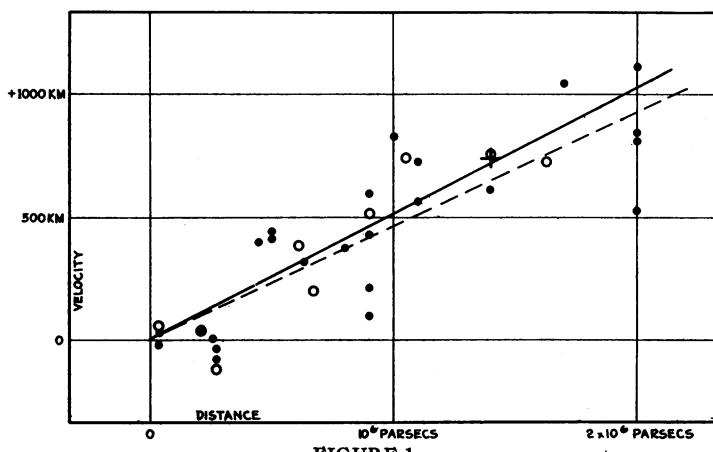


FIGURE 1  
Velocity-Distance Relation among Extra-Galactic Nebulae.

Radial velocities, corrected for solar motion, are plotted against distances estimated from involved stars and mean luminosities of nebulae in a cluster. The black discs and full line represent the solution for solar motion using the nebulae individually; the circles and broken line represent the solution combining the nebulae into groups; the cross represents the mean velocity corresponding to the mean distance of 22 nebulae whose distances could not be estimated individually.

can be excluded, since the observed velocity is so small that the peculiar motion must be large in comparison with the distance effect. The object is not necessarily an exception, however, since a distance can be assigned for which the peculiar motion and the absolute magnitude are both within the range previously determined. The two mean magnitudes, -15.3 and -15.5, the ranges, 4.9 and 5.0 mag., and the frequency distributions are closely similar for these two entirely independent sets of data; and even the slight difference in mean magnitudes can be attributed to the selected, very bright, nebulae in the Virgo Cluster. This entirely unforced agreement supports the validity of the velocity-distance relation in a very

evident matter. Finally, it is worth recording that the frequency distribution of absolute magnitudes in the two tables combined is comparable with those found in the various clusters of nebulae.

The results establish a roughly linear relation between velocities and distances among nebulae for which velocities have been previously published, and the relation appears to dominate the distribution of velocities. In order to investigate the matter on a much larger scale, Mr. Humason at Mount Wilson has initiated a program of determining velocities of the most distant nebulae that can be observed with confidence. These, naturally, are the brightest nebulae in clusters of nebulae. The first definite result,<sup>4</sup>  $v = +3779$  km./sec. for N. G. C. 7619, is thoroughly consistent with the present conclusions. Corrected for the solar motion, this velocity is +3910, which, with  $K = 500$ , corresponds to a distance of  $7.8 \times 10^6$  parsecs. Since the apparent magnitude is 11.8, the absolute magnitude at such a distance is -17.65, which is of the right order for the brightest nebulae in a cluster. A preliminary distance, derived independently from the cluster of which this nebula appears to be a member, is of the order of  $7 \times 10^6$  parsecs.

New data to be expected in the near future may modify the significance of the present investigation or, if confirmatory, will lead to a solution having many times the weight. For this reason it is thought premature to discuss in detail the obvious consequences of the present results. For example, if the solar motion with respect to the clusters represents the rotation of the galactic system, this motion could be subtracted from the results for the nebulae and the remainder would represent the motion of the galactic system with respect to the extra-galactic nebulae.

The outstanding feature, however, is the possibility that the velocity-distance relation may represent the de Sitter effect, and hence that numerical data may be introduced into discussions of the general curvature of space. In the de Sitter cosmology, displacements of the spectra arise from two sources, an apparent slowing down of atomic vibrations and a general tendency of material particles to scatter. The latter involves an acceleration and hence introduces the element of time. The relative importance of these two effects should determine the form of the relation between distances and observed velocities; and in this connection it may be emphasized that the linear relation found in the present discussion is a first approximation representing a restricted range in distance.

<sup>1</sup> *Mt. Wilson Contr.*, No. 324; *Astroph. J., Chicago, Ill.*, **64**, 1926 (321).

<sup>2</sup> *Harvard Coll. Obs. Circ.*, 294, 1926.

<sup>3</sup> *Mon. Not. R. Astr. Soc.*, **85**, 1925 (865-894).

<sup>4</sup> These *PROCEEDINGS*, **15**, 1929 (167).

## Gravitation and the Electron

H. Weyl

Received March 7, 1929

*The Problem.* - The translation of Dirac's theory of the electron into general relativity is not only of formal significance, for, as we know, the Dirac equations applied to an electron in a spherically symmetric electrostatic field yield in addition to the correct energy levels those – or rather the negative of those – of an “electron” with opposite charge but the same mass. In order to do away with these superfluous terms the wave function  $\psi$  must be robbed of one of its pairs  $\psi_1^+, \psi_2^+, \psi_1^-, \psi_2^-$  of components. [1] These two pairs occur unmixed in the action principle except for the term

$$m(\psi_1^+ \bar{\psi}_1^- + \psi_2^+ \bar{\psi}_2^- + \psi_1^- \bar{\psi}_2^+ + \psi_2^- \bar{\psi}_1^+) \quad (1)$$

which contains the mass  $m$  of the electron as a factor. But mass is a gravitational effect: it is the flux of the gravitational field through a surface enclosing the particle in the same sense that charge is the flux of the electric field. [2] In a satisfactory theory it must therefore be as impossible to introduce a non-vanishing mass without the gravitational field as it is to introduce charge without electromagnetic field. It is therefore certain that the term (1) can at most be right in the large scale, but must really be replaced by one which includes gravitation; this may at the same time remove the defects of the present theory.

The direction in which such a modification is to be sought is clear: the field equations arising from an action principle [3] - which shall give the true laws of interaction between electrons, protons and photons only after quantization—contain at present only the Schrödinger–Dirac quantity  $\psi$ ,

which describes the wave field of the *electron*, in addition to the four potentials  $\varphi_p$  of the electromagnetic field. It is unconditionally necessary to introduce the wave field of the *proton* before quantizing. But since the  $\psi$  of the electron can only involve two components,  $\psi_1^+, \psi_2^+$  should be ascribed to the electron and  $\psi_1^-, \psi_2^-$  to the proton. Obviously the present expression,  $-e \cdot \tilde{\psi}, \psi$  for charge-density, [4] being necessarily negative, runs counter to this, and something must consequently be changed in this respect. Instead of one law for the conservation of charge we must have two, expressing the conservation of the number of electrons and protons separately.

If one introduces the quantities  $\frac{e\varphi_p}{ch}$  instead of  $\varphi_p$  (and calls them  $\varphi_p$ ), the field equations contain only the following combinations of atomistic constants: the pure number  $\alpha = e^2/ch$  and  $h/mc$ , the “wave-length” of the electron. Hence the equations certainly do not alone suffice to explain the atomistic behavior of matter with the definite values of  $e$ ,  $m$  and  $h$ . But the subsequent quantization introduces the quantum of action  $h$ , and this together with the wave-length  $h/mc$  will be sufficient, since the velocity of light  $c$  is determined as an absolute measure of velocity by the theory of relativity.

The introduction of the atomic constants by the quantum theory—or at least that of the wave-length—into the field equations has removed the support from under my principle of gauge-invariance, by means of which I had hoped to unify electricity and gravitation. But as I have remarked, [5] it possesses an equivalent in the field equations of quantum theory which is its perfect counterpart in formal respects: the laws are invariant under the simultaneous substitution of  $e^{i\lambda} \cdot \psi$  for  $\psi$  and  $\varphi_p - \frac{\partial\lambda}{\partial x_p}$  for  $\varphi_p$ , where  $\lambda$  is an arbitrary function of position in space and time. The connection of this invariance with the conservation law of electricity remains exactly as before: the fact that the action integral is unaltered by the infinitesimal variation

$$\delta\psi = i\lambda \cdot \psi, \quad \delta\varphi_p = -\frac{\partial\lambda}{\partial x_p}$$

( $\lambda$  an arbitrary infinitesimal function) signifies the identical fulfilment of a dependence between the material and the electromagnetic laws which arise from the action integral by variations of the  $\psi$  and  $\varphi$ , respectively; it means that the conservation of electricity is a double consequence of them, that it follows from the laws of matter as well as electricity. This new principle of gauge invariance, which may go by the same name, has the character of general relativity since it contains an arbitrary function  $\lambda$ , and can certainly only be understood with reference to it.

It was such considerations as these, and not the desire for formal generalizations, which led me to attempt the incorporation of the Dirac theory into the scheme of general relativity. We establish the metric in a world point  $P$  by a “Cartesian” system of axes (instead of the  $g_{pq}$ ) consisting of four vectors  $\mathbf{e}(\alpha)$   $\{\alpha = 0, 1, 2, 3\}$  of which  $\mathbf{e}(1), \mathbf{e}(2), \mathbf{e}(3)$  are real space-like vectors while  $\mathbf{e}(0)/i$  is a real time-like vector of which we expressly demand that it be directed toward the future. A rotation of these axes is an orthogonal or Lorentz transformation which leaves these conditions of reality and sign unaltered. The laws shall remain invariant when the axes in the various points  $P$  are subjected to arbitrary and independent rotations. In addition to these we need four (real) coordinates  $x_p$  ( $p = 0, 1, 2, 3$ ) for the purpose of analytic expression. The components of  $\mathbf{e}(\alpha)$  in this coordinate system are designated by  $e^p(\alpha)$ . We need such local cartesian axes  $\mathbf{e}(\alpha)$  in each point  $P$  in order to be able to describe the quantity  $\psi$  by means of its components  $\psi_1^+, \psi_2^+; \psi_1^-, \psi_2^-$ , for the law of transformation of the components  $\psi$  can only be given for orthogonal transformations as it corresponds to a representation of the orthogonal group which cannot be extended to the group of all linear transformations. The tensor calculus is consequently an unusable instrument for considerations involving the  $\psi$ . [6] In formal aspects our theory resembles the more recent attempts of Einstein to unify electricity and gravitation. [7] But here there is no talk of “distant parallelism”; there is no indication that Nature has availed herself of such an artificial geometry. I am convinced that if there is a physical content in Einstein’s latest formal developments it must come to light in the present connection. It seems to me that it is now hopeless to seek a unification of gravitation and electricity without taking material waves into account.

*Use of the Indices.* – If  $t(\alpha)$  be the components of an arbitrary vector at point  $P$  with respect to the axes  $\mathbf{e}(\alpha)$ , then

$$t^p = \sum_{\alpha} e^p(\alpha) t(\alpha) \quad (2)$$

are its contravariant components in the coordinate system  $x_p$ . Conversely, from the covariant components  $t_p$  referred to the coordinates one obtains the components  $t(\alpha)$  along the axes by the equations

$$t(\alpha) = \sum_p e^p(\alpha) t_p. \quad (3)$$

Equations (2), (3) regulate the transition from one kind of indices to the other (Greek indices referring to the axes, Latin sub- or superscripts

to coordinates.) In the inverse transitions the quantities  $e_p(\alpha)$ , which are defined by

$$e_p(\alpha)e^q(\alpha) = \delta_p^q$$

and which also satisfy

$$e_p(\alpha)e^p(\beta) = \delta(\alpha, \beta)$$

occur as coefficients. The Kronecker  $\delta$  is 1 or 0 according to whether its indices agree or not.

*Symmetry and Conservation of the Energy Density.* – The invariant action

$$\int \mathbf{H} dx \quad (dx = dx_0 dx_1 dx_2 dx_3)$$

contains *matter* (in the extended sense) and gravitation, the first being represented by the  $\psi$  and possibly such additional quantities as the electromagnetic potentials  $\varphi_p$ , the latter by the components  $e^p(\alpha)$  of the  $\mathbf{e}(\alpha)$ . Variation of the first kind of quantities gives rise to the equations of matter, variation of the  $e^p(\alpha)$  to the gravitational equations. We disregard for the present that part of the action which depends only the  $e^p(\alpha)$ , as introduced by Einstein in his classical theory of gravitation (1916), and consider only that part  $\mathbf{H}$  which occurs even in the special theory of relativity. By an arbitrary infinitesimal variation of the  $e^p(\alpha)$  which shall vanish outside of a finite portion of the world an equation

$$\delta \int \mathbf{H} dx = \int \mathbf{t}_p(\alpha) \delta e^p(\alpha) \cdot dx \quad (4)$$

is obtained which defines the components  $\mathbf{t}_p(\alpha)$  of the “energy density.” In consequence of the equations of matter, which are assumed to hold, it is immaterial if or how the quantities describing matter are varied. Because of the invariance of the action (4) must vanish for variations  $\delta e^p(\alpha)$  obtained by 1) subjecting the axes  $\mathbf{e}(\alpha)$  to an infinitesimal rotation which may depend arbitrarily on position and 2) subjecting the coordinates  $x_p$  to an arbitrary infinitesimal transformation, the axes  $\mathbf{e}(\alpha)$  being unaltered. The first process is described by

$$\delta e^p(\alpha) = o(\alpha\beta) e^p(\beta)$$

where  $o(\alpha\beta)$  constitute an anti-symmetric matrix whose elements are arbitrary (infinitesimal) functions of position. This requirement yields the symmetry law:

$$\mathbf{t}_p(\beta) e^p(\alpha) = \mathbf{t}(\alpha, \beta)$$

depends symmetrically on the two indices,  $\alpha$  and  $\beta$ . But it must be observed that this law is not identically fulfilled, as in the old field theory, but

only in consequence of the equations of matter, for if the wave field  $\psi$  be held unchanged the components  $\psi_p$  must undergo a transformation which is induced by the rotation of the axes. If  $\delta x_p = \xi^p$  be the change which the coordinates of point  $P$  undergo in the second process, then the components of the unaltered vector  $\mathbf{e}(\alpha)$  in  $P$  will undergo the change

$$\delta' e^p(\alpha) = \frac{\partial \xi^p}{\partial x_q} \cdot e^q(\alpha).$$

This must, on the other hand, be given by

$$\delta e^p(\alpha) + \frac{\partial e^p(\alpha)}{\partial x_q} \xi^q$$

where  $\delta$  means the difference at two points  $P$  which have the same values  $x_p$ , of coordinates before and after the deformation. From this there arises in the usual way [8]—again assuming the validity of the equations of matter—the differential quasi-conservation law for energy (and linear momentum) whose four components are

$$\frac{\partial \mathbf{t}_p^q}{\partial x_q} + \mathbf{t}_q(\alpha) \frac{\partial e^q(\alpha)}{\partial x_p} = 0. \quad (5)$$

(Only in the special theory of relativity, where the second member is lacking, is it a true conservation law.)

It is not necessary that the integral of  $\mathbf{H}$  be invariant, but only that its variation be. This is the case when  $\mathbf{H}$  differs from a scalar density by a divergence; we then say that the integral is “practically invariant.” Similarly it is only necessary that it be “practically real,” i.e., that the difference between  $\mathbf{H}$  and its complex conjugate be a divergence.

*Gradient of  $\psi$ .* — Let the wave field  $\psi$  be given. The invariant change  $\delta\psi$  of  $\psi$  on going from the point  $P$  to a neighboring point  $P'$  is to be determined as follows. The axes  $\mathbf{e}(\alpha)$  in  $P$  are taken to  $P'$  by parallel displacement:  $\mathbf{e}'(\alpha)$ .  $\psi_p = \psi_p(P)$  being the components of  $\psi$  with respect to the axes  $\mathbf{e}(\alpha)$  at  $P$ , let  $\psi'_p$  be the components of  $\psi$  in  $P'$  relative to this displaced system:  $\delta\psi_p = \psi'_p - \psi_p$ . These  $\delta\psi_p$  depend only on the choice of axes in  $P$  and transform in the same way as the  $\psi_p$  on rotation of these axes. The axes  $\mathbf{e}'(\alpha)$  are obtained from the  $\mathbf{e}(\alpha)$  in  $P'$  by an infinitesimal orthogonal transformation  $d\alpha(\alpha\beta)$ ; consequently the  $\psi'_p$  are obtained from the components  $\psi_p(P')$  by the corresponding linear transformation [9]  $dE$  and we have,  $d\psi_p$  being the differential  $\psi_p(P') - \psi_p(P)$ :

$$\delta\psi = d\psi + dE \cdot \psi.$$

If  $\mathbf{e}(\alpha)$  be taken as the vector  $\overrightarrow{PP'}$  (multiplied by an infinitesimal factor) we write (on ignoring this factor)  $o(\alpha, \beta\gamma)$ ,  $E(\alpha)$ ,  $\psi(\alpha)$  in place of  $do(\beta\gamma)$ ,  $dE$ ,  $d\psi$ :

$$\psi(\alpha) = \left( e^p(\alpha) \frac{\partial}{\partial x_p} + E(\alpha) \right) \psi \quad \text{or} \quad \psi_p = \left( \frac{\partial}{\partial x_p} + E_p \right) \psi.$$

The calculation of  $o$  is accomplished by means of the formula

$$e^p(\gamma) \{ o(\alpha, \beta\gamma) + o(\beta, \gamma\alpha) \} = \frac{\partial e^p(\alpha)}{\partial x_q} e_q(\beta) - \frac{\partial e^p(\beta)}{\partial x_q} e_q(\alpha).$$

The right-hand side of this expression is the “commutator product” of the two vector fields  $\mathbf{e}(\alpha)$  and  $\mathbf{e}(\beta)$ , an invariant (under transformations of the coordinates  $x_p$ ), known from the Lie theory.

*Introduction of Dirac's Action.* — Let  $S(\alpha)$  denote linear transformations which transform  $\psi_1^+, \psi_2^+$ , and  $\psi_1^-, \psi_2^-$  among themselves. They are described by the matrices [10]

$$S(0) = \begin{vmatrix} i & 0 \\ 0 & i \end{vmatrix}, \quad S(1) = \begin{vmatrix} 0 & 1 \\ 1 & 0 \end{vmatrix}, \quad S(2) = \begin{vmatrix} 0 & -i \\ i & 0 \end{vmatrix}, \quad S(3) = \begin{vmatrix} 1 & 0 \\ 0 & -1 \end{vmatrix}$$

for  $\psi_1^+, \psi_2^+$ ; for  $\psi_1^-, \psi_2^-$  the expression for  $S(0)$  is unchanged,  $S(1), S(2)$  and  $S(3)$  assume the opposite sign. The essential fact is that the quantities

$$\tilde{\psi}' S(\alpha) \psi$$

transform like the four components  $t(\alpha)$  of a vector on rotation of the axes  $\mathbf{e}(\alpha), \psi'$  being a quantity of the same kind as  $\psi$ . (In particular,  $\tilde{\psi} S(\alpha) \psi$  is the four-vector flux of probability.) Therefore

$$\tilde{\psi} S(\alpha) \psi(\alpha) = \tilde{\psi} e^p(\alpha) S(\alpha) \frac{\partial \psi}{\partial x_p} + \tilde{\psi} S(\alpha) E(\alpha) \psi$$

is a scalar, and after dividing by the absolute value  $e$  of the determinant

$$|e^p(\alpha)|$$

we obtain a scalar density  $i\mathbf{H}$  whose integral can be employed as action. Division by  $\epsilon$  will be indicated by changing the ordinary letter into the corresponding gothic. The calculation yields

$$\frac{1}{\epsilon} S(\alpha) E(\alpha) = \frac{1}{2} \frac{\partial \mathbf{e}^p(\alpha)}{\partial x_p} S(\alpha) + \frac{1}{2} \mathbf{I}(\alpha) S'(\alpha),$$

where  $S'(\alpha)$  is a transformation analogous to  $S(\alpha)$ : it agrees with  $S(\alpha)$  for  $\psi_1^+, \psi_2^+$ , but is  $-S(\alpha)$  for  $\psi_1^-, \psi_2^-$ .

$$\begin{aligned} e\mathbf{I}(\alpha) &= o(\beta, \gamma\delta) + o(\gamma, \delta\beta) + o(\delta, \beta\gamma) \\ &= \sum \pm \frac{\partial e^p(\beta)}{\partial x_q} e_q(\gamma) e^p(\delta) \end{aligned}$$

where the summation (in addition to that over  $p$  and  $q$ ) is alternating and extends over the six permutations of  $\beta\gamma\delta$  while  $\alpha\beta\gamma\delta$  is an even permutation of the indices 0, 1, 2, 3.

We have yet to investigate whether  $\mathbf{H}$  is practically real. Since

$$S^p = e^p(\alpha)S(\alpha) \quad S(\alpha)E(\alpha)$$

are Hermitean matrices

$$-i\bar{\mathbf{H}} = \frac{1}{\epsilon} \left\{ \frac{\partial \tilde{\psi}}{\partial x_p} e^p(\alpha)S(\alpha)\psi + \tilde{\psi}S(\alpha)E(\alpha)\psi \right\}.$$

The first part is, on neglecting a complete divergence,

$$-\tilde{\psi} \frac{\partial(e^p(\alpha)S(\alpha)\psi)}{\partial x_p} = -\tilde{\psi}e^p(\alpha)S(\alpha) \frac{\partial\psi}{\partial x_p} - \frac{\partial e^p(\alpha)}{\partial x_p} \cdot \tilde{\psi}S(\alpha)\psi.$$

On adding and subtracting  $\mathbf{H}$  and  $\bar{\mathbf{H}}$  we obtain the two action quantities  $\mathbf{m}, \mathbf{m}'$  ( $\mathbf{m}$  = matter):

$$i\mathbf{m} = \tilde{\psi}\mathbf{S}^p \frac{\partial\psi}{\partial x_p} + \frac{1}{2} \frac{\partial e^p(\alpha)}{\partial x_p} \cdot \tilde{\psi}S(\alpha)\psi \quad (6)$$

and

$$\mathbf{m}' = \mathbf{I}(\alpha) \cdot \tilde{\psi}S'(\alpha)\psi. \quad (7)$$

Both are practically, not actually, invariant, the first is practically and the second actually real.

The first is the essential content of the Dirac theory, written in general invariantive form. The corresponding tensor density of energy

$$\mathbf{t}_p(\alpha) = \mathbf{s}_p(\alpha) - e_p(\alpha)\mathbf{s}$$

where

$$\epsilon\mathbf{s}_p(\alpha) = \frac{1}{2i} \left\{ \tilde{\psi}S(\alpha) \frac{\partial\psi}{\partial x_p} - \frac{\partial\tilde{\psi}}{\partial x_p} S(\alpha)\psi \right\}$$

and  $\mathbf{s}$  the contracted

$$\mathbf{s}_p(\alpha) e^p(\alpha).$$

It has already been given in the literature for the Dirac theory (special relativity). For the electron of hydrogen in the normal state we find that the integral

$$\int \mathbf{t}_0^0 dx_1 dx_2 dx_3$$

extended over a section  $x_0 = t = \text{const.}$ , which should yield the mass, has the value  $m/\sqrt{1 - \alpha^2}$  ( $\alpha$  the fine structure constant); this is a reasonable result, since  $m$  is to be taken as proper mass and in the Bohr theory  $\alpha c$  is the velocity of the electron in the normal state.

It is worthy of note that there occurs in addition the action  $\mathbf{m}'$ , which is unknown to special relativity since it vanishes for constant  $e^p(\alpha)$ .

*The Electromagnetic Field.* – In the Dirac theory the influence of an electromagnetic field is taken into account by replacing the operator  $\frac{\partial}{\partial x_p}$  affecting the  $\psi$  by  $\frac{\partial}{\partial x_p} + i\varphi_p$ . This yields an additional term of the form  $i\varphi(\alpha) \cdot \tilde{\psi} S(\alpha) \psi$  in the action. On comparing this with (6) and (7) one might think that  $\epsilon \cdot \varphi(\alpha)$  is to be identified with  $\partial \mathbf{e}^p(\alpha)/\partial x_p$ , or  $\mathbf{I}(\alpha)$ . Disregarding the material waves  $\psi$  one would then have a theory of electricity of the same kind as the latest development of Einstein; the  $\varphi$  are expressed in terms of the  $e^p(\alpha)$  in a way which is invariant under transformations of the coordinates, but only permit the cogredient rotations of the axes in all points  $P$  (distant parallelism). However, I believe one must proceed otherwise in order to bring in the electromagnetic field. We have previously not mentioned the fact that the linear transformation of the  $\psi$  corresponding to a rotation of axes is not uniquely determined. It is indeed possible to normalize this transformation, and we have tacitly based our calculations on such a normalization; but the normalizing is itself a double-valued process and this reveals its artificial character. The mathematical connection is this: we are to find a linear transformation of the four components of  $\psi$  such that the quantities

$$\tilde{\psi} S(\alpha) \psi \tag{8}$$

suffer the given rotation. [11] Obviously it remains unaltered when  $\psi_1^+, \psi_2^+$  are multiplied by  $e^{i\lambda^+}, \psi_1^-, \psi_2^-$  by  $e^{i\lambda^-}$  (transformation  $L$ ) where  $\lambda^+$  and  $\lambda^-$  are arbitrary real numbers. It is readily seen that the transformations  $L$  are the only ones which induce the identical transformation of the quantities (8); i.e., which satisfy the four equations

$$\tilde{L} S(\alpha) L = S(\alpha).$$

A transformation of the four components of  $\psi$  which multiplies  $\psi_1^+, \psi_2^+$  by a number  $a^+$  and  $\psi_1^-, \psi_2^-$  by a number  $a^-$  will be called *spinless quantity*  $A = [a^+, a^-]$ .

In consequence of this the  $dE$  employed above is only determined to within the addition of an arbitrary spinless imaginary quantity  $i dF$ . Hence we obtain an additive term

$$iF(\alpha) \cdot \tilde{\psi} \mathbf{S}(\alpha) \psi$$

in the action  $\mathbf{m}$ , in which the  $F(\alpha)$  are real spinless quantities  $[\varphi^+(\alpha), \varphi^-(\alpha)]$  constituting the components of a vector with respect to the axes.  $\partial/\partial x_p$  is to be replaced by  $\partial/\partial x_p + iF_p$ . We now employ the letter  $\mathbf{m}$  to denote this completed action. The introduction of  $F$  obviously brings with it invariance under the simultaneous replacement of

$$\psi \text{ by } e^{iL} \cdot \psi \quad \text{and} \quad F_p \text{ by } F_p - \frac{\partial L}{\partial x_p}$$

where  $L$  is a real spinless quantity which depends arbitrarily on position. This “gauge invariance” shows why it is impossible to employ the scalar (1) as an action function.

We must naturally interpret  $F_p$  as the components of electromagnetic potential. The two fields  $\varphi_p^+, \varphi_p^-$ , are independent of the choice of the axes  $\mathbf{e}(\alpha)$  except that they are interchanged on transition from right- to a left-handed system of axes.

$$\frac{\partial \varphi_q^+}{\partial x_p} - \frac{\partial \varphi_p^+}{\partial x_q} = \varphi_{pq}^+ \quad \text{and the corresponding} \quad \varphi_{pq}^-$$

are gauge-invariant anti-symmetric tensors.

In analogy to the Maxwellian action quantity, and in accordance with the above properties, the two functions  $\mathbf{f}, \mathbf{f}'$  ( $\mathbf{f}$  = electromagnetic field) defined by

$$\epsilon \mathbf{f} = \varphi_{pq}^+ \varphi_-^{pq} \quad \text{and} \quad \epsilon \mathbf{f}' = \varphi_{pq}^+ \varphi_+^{pq} + \varphi_{pq}^- \varphi_-^{pq} \quad (9)$$

are to be considered in choosing the scalar density of electromagnetic action. The identification of  $F$  with the electromagnetic potential is then justified by the fact that the entity which is represented by  $F$  influences and is influenced by matter in exactly the same way as the electromagnetic potentials. *If our view is correct, then the electromagnetic field is a necessary accompaniment of the matter-wave field and not of gravitation.*

If the action, insofar as its dependence on the  $\psi$  and  $\varphi$  is concerned, is an additive combination of  $\mathbf{m}, \mathbf{m}', \mathbf{f}, \mathbf{f}'$  we obtain the two conservation theorems

$$\frac{\partial \rho_+^p}{\partial x_p} = 0 \quad \text{and} \quad \frac{\partial \rho_-^p}{\partial x_p} = 0 \quad (10)$$

where

$$\rho_+^p = \tilde{\psi}^+ \mathbf{S}^p \psi^+$$

contains only  $\psi_1^+$  and  $\psi_2^+$  and similarly for  $\rho_-^p$ . They are a double consequence of the field laws, that is, of the equations of matter in the narrow sense as well as of the electromagnetic equations. These two identities obtaining thus between the field equations are an immediate consequence of the gauge invariance. In consequence of (10) the two integrals  $n^+$  and  $n^-$  of  $\rho = \rho^0$ :

$$n = \int \rho dx_1 dx_2 dx_3$$

which are to be extended over a section  $x_0 = \text{const.}$  are invariants which are independent of the “time”  $x_0$ . [12] Since they are interchanged on transition from right- to left-handed axes their values, which are absolute constants of nature, must be equal if both kinds of axes are to be equally permissible. We normalize them, in accordance with the interpretation of  $\psi$  as probability, by

$$n^+ = 1, \quad n^- = 1. \quad (11)$$

We have already mentioned how the quasi-conservation laws for energy, momentum and moment- of momentum, are related to invariance under transformation of coordinates and rotation of axes  $\mathbf{e}(\alpha)$ .

A stationary solution is characterized by the fact that  $e^p(\alpha)$  and  $F(\alpha)$  are independent of time  $x_0 = t$ , while the  $\psi^+$  contain the time in an exponential factor  $e^{i\nu t}$ , the  $\psi^-$  in a factor  $e^{i\nu' t}$ ;  $\nu$  and  $\nu'$  need not be equal.

*Gravitation.* – We consider as the gravitational part of the action the practically (not actually) invariant density  $\mathbf{g}$  ( $\mathbf{g}$  = gravitation) which underlies Einstein’s “classical” theory of gravitation and which depends only on the  $e^p(\alpha)$  and their first derivatives. It is most appropriate to carry through anew the entire calculation, which leads through the Riemann curvature tensor, in terms of the  $e^p(\alpha)$ ; we find

$$\epsilon \mathbf{g} = o(\alpha, \alpha\gamma)o(\beta, \beta\gamma) + o(\alpha, \beta\gamma)o(\beta, \alpha\gamma).$$

The “cosmological term”  $\mathbf{g}'$  is given by  $\epsilon \mathbf{g}' = 1$ .

If gravitation be represented by  $\mathbf{g}$ , one can, as is well known, add to the material + electromagnetic energy  $\mathbf{t}_p^q$  a gravitational energy  $\mathbf{v}_p^q$  in such a way that the sum satisfies a true conservation law. [13] Designating the total differential of  $\mathbf{g}$  considered as a function of  $e^p(\alpha)$  and  $e_q^p(\alpha) = \frac{\partial e^p(\alpha)}{\partial x_q}$  by

$$\delta\mathbf{g} = \mathbf{g}_p(\alpha)\delta e^p(\alpha) + \mathbf{g}_p^q(\alpha)\delta e_q^p(\alpha),$$

then

$$-\mathbf{v}_p^q = \mathbf{g}_\tau^q(\alpha) \frac{\partial e^\tau(\alpha)}{\partial x_p} - \delta_p^q \mathbf{g}.$$

We obtain thus an invariant constant mass  $m$  which must be one of the characteristic universal constants of Nature.

*Doubts, Prospects.* – (11) is to be interpreted as the law of conservation of the number (or charge) of electrons and protons. Therefore we ascribe  $\psi^+$  and  $\psi^-$  to the electron and to the proton, respectively. Taking  $\mathbf{f}$  as the electromagnetic part of the action, which seems plausible to me, we then obtain Maxwell's equations in the sense that the proton generates the field  $\varphi^+$  and the electron  $\varphi^-$ ; whereas in accordance with the equations of matter  $\varphi^+$  will effect only the electron and  $\varphi^-$  the proton. This is not as obtruse as it may sound; on the contrary, the previous theory leads to entirely false results if the potential due to the electron, which at large distances neutralizes that due to the nucleus, reacts on the electron itself, as Schrödinger has pointed out with emphasis. [14] It may indeed seem queer that  $\psi^+$  and  $\psi^-$  are here equally permissible, since we know that positive and negative electricity are fundamentally different—that protons and electrons have different mass. But if we neglect the gravitational and electrical energy in comparison with the material the mass  $m$  falls into two parts  $m^+$  and  $m^-$  which are, however, not strictly constant. It is possible that  $m^+$  and  $m^-$  are different if our equations admit two classes of solutions which are interchanged on transition from right- to left-handed axes – as in the Dirac theory, the spherically symmetric hydrogen problem admits several solutions for the normal state which are not themselves spherically symmetric but which are transformed among themselves by rotation.

As far as we know  $\mathbf{m}, \mathbf{g}$  and one of the two quantities  $\mathbf{f}$  or  $\mathbf{f}'$  are indispensable for the explanation of the phenomena. I am inclined to believe that the action is composed additively of  $\mathbf{m}, \mathbf{f}$  and  $\mathbf{g}$ .

It should be noted that our field equations contain neither the theory of a single electron nor that of a single proton. One might rather consider them as me laws governing a hydrogen atom consisting of an electron and a

proton; but here again, the problem of interaction between the two may first require quantization. What we have obtained is solely a field scheme which can only be applied to and compared with experience after the quantization has been accomplished. We know from the Pauli exclusion principle [15] what commutation rules are to be applied in the quantization of  $\psi^+$ ; those for  $\psi^-$  must be the same in our theory. The commutation relations between  $\psi^+$  and  $\psi^-$  are as yet entirely unknown. Those of the electromagnetic field (photons) are almost completely known. In this respect we know nothing concerning the gravitational field. The commutation rules for  $F$  are here almost completely fixed by those for  $\psi$  by the condition that these latter be unaltered when  $\psi$  is given the increment  $\delta\psi = iF(\alpha)\psi$ . That the rules thus obtained are in agreement with experience is indeed a support for our theory; i.e., it tells us why the “anti-symmetric,” Pauli–Fermi statistics; for electrons leads to the “symmetric” Bose–Einstein statistics for photons. A definite decision can, however, first be reached when the barrier which hems the progress of quantum theory is overcome: the quantization of the field equations.

## References

- [1] I employ the same notation as in my book *Gruppentheorie und Quantenmechanik*, Leipzig, 1928 (cited as GQ) except that I here write  $\psi_1^+, \psi_2^+, \psi_1^-, \psi_2^-$  in place of  $\psi_1\psi_2, \psi_3\psi_4$ . Cf. in particular § 25, 39, 44.  
 $\frac{\hbar}{2\pi}$  is Planck’s constant.
- [2] Cf. H. Weyl, *Space–Time–Matter*, London, 1922 (cited as S T M), § 33.
- [3] G Q, pp. 199, 200.
- [4] The circumflex indicates transition to the conjugate of the transposed matrix (Hermitean conjugate). The four components of  $\psi$  are considered as the elements of a matrix with four rows and one column.
- [5] G Q, p. 88.
- [6] Attempts to employ only the tensor calculus have been made by Tetrode (Z. Physik, 50, 336 (1928)); J. M. Whittaker (*Proc. Cambr. Phil. Soc.*, 25, 501 (1928)), and others; I consider them misleading.

- [7] *Sitzungsber. Berl. Akad.*, **1928**, pp. 217, 224; 1929.
- [8] Cf. the analogous considerations in *G T M*, pp. 233–237.
- [9] Capital Latin letters (except  $P$  for point) denote linear transformations of the four components of  $\psi$ .
- [10] In *G Q*, loc. cit., they are denoted by  $s'_\alpha$ .
- [11] It is to be borne in mind that under the influence of a proper Lorentz transformation the  $\psi^+$  components – as well as the  $\psi^-$  – are transformed among themselves. Only when the improper operations of the Lorentz group, the reflection

$$\mathbf{e}(0) \rightarrow \mathbf{e}(0), \quad \mathbf{e}(\alpha) \rightarrow -\mathbf{e}(\alpha) \quad [\alpha = 1, 2, 3],$$

is taken into account is it necessary to use both pairs of components together.

- [12] *S T M*, p. 289.
- [13] The derivation in *S T M*, pp. 269, 270, can be adapted to the new analytic formulation of the gravitational field.
- [14] E. Schrödinger, Ann. Physik, **82**, 265-272 (1927); in particular p. 270.
- [15] P. Jordan and E. Wigner, *Z. Physik*, **47**, 1928, 631; *G Q*, § 44.

*Correction made on the proofs (March 4, (1929)).* – The calculation of the action density  $m$  contains an error which should be corrected as follows. The spacial components  $\alpha = 1, 2, 3$  of  $l(\alpha)$  are pure imaginary and the temporal component  $l(0)$  real, not the opposite as I had assumed. In the definition of  $m'$  we must therefore divide the right-hand side by  $2i$ . But this has as consequence that the  $\mathbf{H} = m + m'$  obtained from the calculation is practically real and not composed of a real and an imaginary part. We therefore obtain but one invariant action density for matter:  $m + m'$ . To the tensor density of energy arising from  $m$  must naturally be added the term arising from  $m'$ .



## The Production of High Speed Protons Without the use of High Voltages

E.O. Lawrence

(Received 1931)

A method for the production of high speed protons without the use of high voltages was described before the meeting of the National Academy of Sciences last September (Lawrence and Edlefson, Science **72**, 376-377, 1930). Later before the American Physical Society (Lawrence and Livingston, Phys. Rev. **37**, 1707, 1931) results of a preliminary study of the practicability of this method were presented. In this preliminary experimental work 80,000-volt hydrogen molecule ions were successfully produced in a vacuum tube in which the maximum applied potential was less than 2,000 volts, and the conclusion of the experiments was that there are no serious difficulties in the way of producing 1,000,000-volts protons in this indirect manner.

The important conclusion has now been confirmed. A magnet having pole faces nine inches in diameter and producing a field of 15,000 gauss has recently been constructed and with its aid protons and hydrogen molecule ions having energies in excess of one half million volt-electrons have been produced.

The magnitudes of the high speed hydrogen ion currents turned out to be surprisingly large, being in excess of one-tenth of one microampere. The proton currents were about one-tenth this value.

The voltage amplification obtained in the present experiment was approximately one hundred. That is to say, about five thousand volts were

applied to the tube for the production of five hundred thousand volt ions. This amplification was limited by the slit system used to select out the high speed ions, and can be greatly increased by better design of this part of the tube.

There can be little doubt that one million volt ions will be produced with intensities as great as here recorded when the present experimental tube is enlarged to make full use of the magnet. This alteration is now being carried.

These experiments make it evident that with quite ordinary laboratory facilities proton beams having great enough energies for nuclear studies can be readily produced with intensities far exceeding the intensities of beams of alpha-particles from radioactive sources.

Possible the most interesting consequence of these experiments is that it appears now that the production of 10,000,000-volt protons can be readily accomplished when a suitably larger magnet and high frequency oscillator are available. The importance of the production of protons of such speeds can hardly be overestimated and it is our hope that the necessary equipment for doing this will be made available to us.

We are very much indebted to the Federal Telegraph Company, through the courtesy of Dr. Leonard F. Fuller, Vice-president, for the loan of essential parts of the apparatus used in this work.

Ernest O. Lawrence  
M. Stanley Livingston

University of California,  
July 20, 1931.

## Quantised Singularities in the Electromagnetic Field

P.A.M. Dirac

Received May 29, 1931

### § 1. *Introduction*

The steady progress of physics requires for its theoretical formulation a mathematics that gets continually more advanced. This is only natural and to be expected. What, however, was not expected by the scientific workers of the last century was the particular form that the line of advancement of the mathematics would take, namely, it was expected that the mathematics would get more and more complicated, but would rest on a permanent basis of axioms and definitions, while actually the modern physical developments have required a mathematics that continually shifts its foundations and gets more abstract. Non-euclidean geometry and non-commutative algebra, which were at one time considered to be purely fictions of the mind and pastimes for logical thinkers, have now been found to be very necessary for the description of general facts of the physical world. It seems likely that this process of increasing abstraction will continue in the future and that advance in physics is to be associated with a continual modification and generalisation of the axioms at the base of the mathematics rather than with a logical development of any one mathematical scheme on a fixed foundation.

There are at present fundamental problems in theoretical physics awaiting solution, e.g., the relativistic formulation of quantum mechanics and the nature of atomic nuclei (to be followed by more difficult ones such as the problem of life), the solution of which problems will presumably require a more drastic revision of our fundamental concepts than any that have gone

before. Quite likely these changes will be so great that it will be beyond the power of human intelligence to get the necessary new ideas by direct attempts to formulate the experimental data in mathematical terms. The theoretical worker in the future will therefore have to proceed in a more indirect way. The most powerful method of advance that can be suggested at present is to employ all the resources of pure mathematics in attempts to perfect and generalise the mathematical formalism that forms the existing basis of theoretical physics, and after each success in this direction, to try to interpret the new mathematical features in terms of physical entities (by a process like Eddington's Principle of Identification).

A recent paper by the author<sup>1</sup> may possibly be regarded as a small step according to this general scheme of advance. The mathematical formalism at that time involved a serious difficulty through its prediction of negative kinetic energy values for an electron. It was proposed to get over this difficulty, making use of Fault's Exclusion Principle which does not allow more than one electron in any state, by saying that in the physical world almost all the negative-energy states are already occupied, so that our ordinary electrons of positive energy cannot fall into them. The question then arises to the physical interpretation of the negative-energy states, which on this view really exist. We should expect the uniformly filled distribution of negative-energy states to be completely unobservable to us, but an unoccupied one of these states, being something exceptional, should make its presence felt as a kind of hole. It was shown that one of these holes would appear to us as a particle with a positive energy and a positive charge and it was suggested that this particle should be identified with a proton. Subsequent investigations, however, have shown that this particle necessarily has the same mass as an electron<sup>2</sup> and also that, if it collides with an electron, the two will have a chance of annihilating one another much too great to be consistent with the known stability of matter.<sup>3</sup>

It thus appears that we must abandon the identification of the holes with protons and must find some other interpretation for them. Following Oppenheimer,<sup>4</sup> we can assume that in the world as we know it, *all*, and not merely nearly all, of the negative-energy states for electrons are occupied. A hole, if there were one, would be a new kind of particle, unknown to experimental physics, having the same mass and opposite charge to an electron.

<sup>1</sup>Proc. Roy. Soc.,' A, vol. 126, p. 360 (1930).

<sup>2</sup>H. Weyl, 'Gruppentheorie and Quantenmechanik,' 2nd ed. p. 234 (1931).

<sup>3</sup>I. Tamm, 'Z. Physik,' vol. 62, p. 545 (1930); J. B. Oppenheimer, 'Phys. Rev.,' vol. 35, p. 939 (1930); P. Dirac, 'Proc. Camb. Philos. Soc.,' vol. 26, p. 361 (1930).

<sup>4</sup>J. R. Oppenheimer, 'Phys. Rev.,' vol. 35, p. 562 (1930).

We may call such a particle an anti-electron. We should not expect to find any of them in nature, on account of their rapid rate of recombination with electrons, but if they could be produced experimentally in high vacuum they would be quite stable and amenable to observation. An encounter between two hard  $\gamma$ -rays (of energy at least half a million volts) could lead to the creation simultaneously of an electron and anti-electron, the probability of occurrence of this process being of the same order of magnitude as that of the collision of the two  $\gamma$ -rays on the assumption that they are spheres of the same size as classical electrons. This probability is negligible, however, with the intensities of  $\gamma$ -rays at present available.

The protons on the above view are quite unconnected with electrons. Presumably the protons will have their own negative-energy states, all of which normally are occupied, an unoccupied one appearing as an anti-proton. Theory at present is quite unable to suggest a reason why there should be any differences between electrons and protons.

The object of the present paper is to put forward a new idea which is in many respects comparable with this one about negative energies. It will be concerned essentially, not with electrons and protons, but with the reason for the existence of a smallest electric charge. This smallest charge is known to exist experimentally and to have the value  $e$  given approximately by<sup>5</sup>

$$hc/e^2 = 137. \quad (1)$$

The theory of this paper, while it looks at first as though it will give a theoretical value for  $e$ , is found when worked out to give a connection between the smallest electric charge and the smallest magnetic pole. It shows, in fact, a symmetry between electricity and magnetism quite foreign to current views. It does not, however, force a complete symmetry, analogous to the fact that the symmetry between electrons and protons is not forced when we adopt Oppenheimer's interpretation. Without this symmetry, the ratio on the left-hand aide of (1) remains, from the theoretical standpoint, completely undetermined and if we insert the experimental value 137 in our theory, it introduces quantitative differences between electricity and magnetism so large that one can understand why their qualitative similarities have not been discovered experimentally up to the present.

---

<sup>5</sup> $h$  means Planck's divided by  $2\pi$ .

## § 2. Non-integrable Phases for Wave Functions.

We consider a particle whose motion is represented by a wave function  $\psi$  which is a function of  $x, y, z$  and  $t$ . The precise form of the wave equation and whether it is relativistic or not, are not important for the present theory. We express  $\psi$  in the form

$$\psi = Ae^{i\gamma}, \quad (2)$$

where  $A$  and  $\gamma$  are real functions of  $x, y, z$  and  $t$ , denoting the amplitude and phase of the wave function. For a given state of motion of the particle,  $\psi$  will be determined except for an arbitrary constant numerical coefficient, which must be of modulus unity if we impose the condition that shall be normalised.

The indeterminacy in  $\psi$  then consists in the possible addition of an arbitrary constant to the phase  $\gamma$ . Thus the value of  $\gamma$  at a particular point has no physical meaning and only the difference between the values of  $\gamma$  at two different points is of any importance.

This immediately suggests a generalisation of the formalism. We may assume that  $\gamma$  has no definite value at a particular point, but only a definite difference in values for any two points. We may go further and assume that this difference is not definite unless the two points are neighbouring. For two distant points there will then be a definite phase difference only relative to some curve joining them and different curves will in general give different phase differences. The total change in phase when one goes round a closed curve need not vanish.

Let us examine the conditions necessary for this non-integrability of phase not to give rise to ambiguity in the applications of the theory. If we multiply  $\psi$  by its conjugate complex  $\phi$  we get the density function, which has a direct physical meaning. This density is independent of the phase of the wave function, so that no trouble will be caused in this connection by any indeterminacy of phase. There are other more general kinds of applications, however, which must also be considered. If we take two different wave functions  $\psi_m$  and  $\psi_n$  we may have to make use of the product  $\phi_m\psi_n$ . The integral

$$\int \phi_m\psi_n dx dy dz$$

is a number, the square of whose modulus has a physical meaning, namely, the probability of agreement of the two states. In order that the integral may have a definite modulus the integrand, although it need not have a definite phase at each point, must have a definite phase difference between any two points, whether neighbouring or not. Thus the change in phase in

$\phi_m \psi_n$  round a closed curve must vanish. This requires that the change in phase in  $\psi_n$  round a closed curve shall be equal and opposite to that  $\phi_m$  and hence the same as that in  $\psi_m$ . We thus get the general result: *The change in phase of a wave function round any closed curve must be the same for all the wave functions.*

It can easily be seen that this condition, when extended so as to give the same uncertainty of phase for transformation functions and matrices representing observables (referring to representations in which  $x, y$  and  $z$  are diagonal) as for wave functions, is sufficient to insure that the non-integrability of phase gives rise to no ambiguity in all applications of the theory. Whenever a  $\psi_n$  appears, if it is not multiplied into a  $\phi_m$ , it will at any rate be multiplied into something of a similar nature to a  $\phi_m$ , which will result in the uncertainty of phase cancelling out, except for a constant which does not matter. For example, if  $\psi_n$  is to be transformed to another representation in which, say, the observables  $\xi$ , are diagonal, it must be multiplied by the transformation function  $(\xi, xyzt)$  and integrated with respect to  $x, y$  and  $z$ . This transformation function will have the same uncertainty of phase as a  $\phi$ , so that the transformed wave function will have its phase determinate, except for a constant independent of  $\xi$ . Again, if we multiply  $\psi_n$  by a matrix  $(x'y'z't|\alpha|x''y''z''t)$ , representing an observable  $\alpha$ , the uncertainty in the phase as concerns the column [specified by  $x'', y'', z'', t$ ] will cancel the uncertainty in  $\psi_n$  and the uncertainty as concerns the row will survive and give the necessary uncertainty in the new wave function  $\alpha\psi_n$ . The superposition principle for wave functions will be discussed a little later and when this point is settled it will complete the proof that all the general operations of quantum mechanics can be carried through exactly as though there were no uncertainty in the phase at all.

The above result that the change in phase round a closed curve must be the same for all wave functions means that this change in phase must be something determined by the dynamical system itself (and perhaps also partly by the representation) and must be independent of which state of the system is considered. As our dynamical system is merely a simple particle, it appears that the non-integrability of phase must be connected with the field of force in which the particle moves.

For the mathematical treatment of the question we express  $\psi$ , more generally than (2), as a product

$$\psi = \psi_1 e^{i\beta}, \quad (3)$$

where  $\psi_1$  is any ordinary wave function (i.e., one with a definite phase at each point) whose modulus is everywhere equal to the modulus of  $\psi$ . The

uncertainty of phase is thus put in the factor  $e^{i\beta}$ . This requires that  $\beta$  shall not be a function of  $x, y, z, t$  having a definite value at each point, but  $\beta$  must have definite derivatives

$$\kappa_x = \frac{\partial \beta}{\partial x}, \quad \kappa_y = \frac{\partial \beta}{\partial y}, \quad \kappa_z = \frac{\partial \beta}{\partial z}, \quad \kappa_0 = \frac{\partial \beta}{\partial t},$$

at each point, which do not in general satisfy the conditions of integrability  $\partial \kappa_x / \partial y = \partial \kappa_y / \partial x$ , etc. The change in phase round a closed curve will now be, by Stokes' theorem,

$$\int (\kappa, \mathbf{ds}) = \int (\text{curl } \kappa, \mathbf{dS}), \quad (4)$$

where  $ds$  (a 4-vector) is an element of arc of the closed curve and  $dS$  (a 6-vector) is an element of a two-dimensional surface whose boundary is the closed curve. The factor  $\psi_1$  does not enter at all into this change in phase.

It now becomes clear that the non-integrability of phase is quite consistent with the principle of superposition, or, stated more explicitly, that if we take two wave functions  $\psi_m$  and  $\psi_n$  both having the same change in phase round any closed curve, any linear combination of them  $c_m \psi_m + c_n \psi_n$  must also have this same change in phase round every closed curve. This is because  $\psi_m$  and  $\psi_n$  will both be expressible in the form (3) with the same factor  $e^{i\beta}$  (i.e., the same  $\kappa$ 's) but different  $\psi_1$ 's, so that the linear combination will be expressible in this form with the same  $e^{i\beta}$  again, and this  $e^{i\beta}$  determines the change in phase round any closed curve. We may use the same factor  $e^{i\beta}$  in (3) for dealing with all the wave functions of the system, but we are not obliged to do so, since only  $\text{curl } \kappa$  is fixed and we may use  $\kappa$ 's differing from one another by the gradient of a scalar for treating the different wave functions.

From (3) we obtain

$$-ih \frac{\partial}{\partial x} \psi = e^{i\beta} \left( -ih \frac{\partial}{\partial x} + h\kappa_x \right) \psi_1, \quad (5)$$

with similar relations for the  $y, z$  and  $t$  derivatives. It follows that if  $\psi$  satisfies any wave equation, involving the momentum and energy operators  $\mathbf{p}$  and  $W$ ,  $\psi_1$  will satisfy the corresponding wave equation in which  $\mathbf{p}$  and  $W$  have been replaced by  $\mathbf{p} + h\kappa$  and  $W - h\kappa_0$  respectively.

Let us assume that  $\psi$  satisfies the usual wave equation for a free particle in the absence of any field. Then  $\psi_1$  will satisfy the usual wave equation for a particle with charge  $-e$  moving in an electromagnetic field whose potentials are

$$\mathbf{A} = hc/e \cdot \kappa, \quad \mathbf{A}_0 = -h/e \cdot \kappa_0. \quad (6)$$

Thus, since  $\psi_1$  is just an ordinary wave function with a definite phase, our theory reverts to the usual one for the motion of an electron in an electromagnetic field. This gives a physical meaning to our non-integrability of phase. We see that we must have the wave function  $\psi$  always satisfying the same wave equation, whether there is a field or not, and the whole effect of the field when there is one is in making the phase non-integrable.

The components of the 6-vector  $\text{curl } \kappa$  appearing in (4) are, apart from numerical coefficients, equal to the components of the electric and magnetic fields  $\mathbf{E}$  and  $\mathbf{H}$ . They are, written in three-dimensional vector-notation,

$$\text{curl } \kappa = \frac{e}{hc} \mathbf{H}, \quad \text{grad } \kappa_0 - \frac{\partial \kappa}{\partial t} = \frac{e}{h} \mathbf{E}. \quad (7)$$

The connection between non-integrability of phase and the electromagnetic field given in this section is not new, being essentially just Weyl's Principle of Gauge Invariance in its modern form.<sup>6</sup> It is also contained in the work of Iwanenko and Fock,<sup>7</sup> who consider a more general kind of non-integrability based on a general theory of parallel displacement of half-vectors. The present treatment is given in order to emphasise that non-integrable phases are perfectly compatible with all the general principles of quantum mechanics and do not in any way restrict their physical interpretation.

### § 3. Nodal Singularities.

We have seen in the preceding section how the non-integrable derivatives  $\kappa$  of the phase of the wave function receive a natural interpretation in terms of the potentials of the electromagnetic field, as the result of which our theory becomes mathematically equivalent to the usual one for the motion of an electron in an electromagnetic field and gives us nothing new. There is, however, one further fact which must now be taken into account, namely, that a phase is always undetermined to the extent of an arbitrary integral multiple of  $2\pi$ . This requires a reconsideration of the connection between the  $\kappa$ 's and the potentials and leads to a new physical phenomenon.

The condition for an unambiguous physical interpretation of the theory was that the change in phase round a closed curve should be the same for all

---

<sup>6</sup>H. Weyl, 'Z. Physik,' vol. 56, p. 330 (1929).

<sup>7</sup>D. Iwanenko and V. Fock, 'C. R.,' vol. 188, p. 1470 (1929); V. Fock, 'Z. Physik.' vol. 57, p. 261 (1929). The more general kind of non-integrability considered by these authors does not seem to have any physical application.

wave functions. This change was then interpreted, by equations (4) and (7), as equal to (apart from numerical factors) the total flux: through the closed curve of the 6-vector  $\mathbf{E}, \mathbf{H}$  describing the electromagnetic field. Evidently these conditions must now be relaxed. The change in phase round a closed curve may be different for different wave functions by arbitrary multiples of  $2\pi$  and is thus not sufficiently definite to be interpreted immediately in terms of the electromagnetic field.

To examine this question, let us consider first a very small closed curve. Now the wave equation requires the wave function to be continuous (except in very special circumstances which can be disregarded here) and hence the change in phase round a small closed curve must be small. Thus this change cannot now be different by multiples of  $2\pi$  for different wave functions. It must have one definite value and may therefore be interpreted without ambiguity in terms of the flux of the 6-vector  $E, H$  through the small closed curve, which flux must also be small.

There is an exceptional case, however, occurring when the wave function vanishes, since then its phase does not have a meaning. As the wave function is complex, its vanishing will require two conditions, so that in general the points at which it vanishes will lie along a line.<sup>8</sup> We call such a line a nodal line. If we now take a wave function having a nodal line passing through our small closed curve, considerations of continuity will no longer enable us to infer that the change in phase round the small closed curve must be small. All we shall be able to say is that the change in phase will be close to  $2\pi n$  where  $n$  is some integer, positive or negative. This integer will be a characteristic of the nodal line. Its sign will be associated with a direction encircling the nodal line, which in turn may be associated with a direction along the nodal line.

The difference between the change in phase round the small closed curve and the nearest  $2\pi n$  must now be the same as the change in phase round the closed curve for a wave function with no nodal line through it. It is therefore this difference that must be interpreted in terms of the flux of the 6-vector  $\mathbf{E}, \mathbf{H}$  through the closed curve. For a closed curve in three-dimensional space, only magnetic flux will come into play and hence we obtain for the change in phase round the small closed curve

$$2\pi n + e/hc \cdot \int (\mathbf{H}, \mathbf{dS}).$$

---

<sup>8</sup>We are here considering, for simplicity in explanation, that the wave function is in three dimensions. The passage to four dimensions makes no essential change in the theory. The nodal lines then become two-dimensional nodal surfaces, which can be encircled by curves in the same way as lines are in three dimensions.

We can now treat a large closed curve by dividing it up into a network of small closed curves lying in a surface whose boundary is the large closed curve. The total change in phase round the large closed curve will equal the sum of all the changes round the small closed curves and will therefore be

$$2\pi \sum n + e/hc \cdot \int (\mathbf{H}, d\mathbf{S}), \quad (8)$$

the integration being taken over the surface and the summation over all nodal lines that pass through it, the proper sign being given to each term in the sum. This expression consists of two parts, a part  $e/hc \cdot \int (\mathbf{H}, d\mathbf{S})$  which must be the same for all wave functions and a part  $2\pi \sum n$  which may be different for different wave functions.

Expression (8) applied to any surface is equal to the change in phase round the boundary of the surface. Hence expression (8) applied to a closed surface must vanish. It follows that  $\sum n$ , summed for all nodal lines crossing a closed surface, must be the same for all wave functions and must equal  $-e/2\pi hc$  times the total magnetic flux crossing the surface.

If  $\sum n$  does not vanish, some nodal lines must have end points inside the closed surface, since a nodal line without such end point must cross the surface twice (at least) and will contribute equal and opposite amounts to  $\sum n$  at the two points of crossing. The value of  $\sum n$  for the closed surface will thus equal the sum of the values of  $n$  for all nodal lines having end points inside the surface. This sum must be the same for all wave functions. Since this result applies to any closed surface, it follows that *the end points of nodal lines must be the same for all wave functions. These end points are then points of singularity in the electromagnetic field.* The total flux of magnetic field crossing a small closed surface surrounding one of these points is

$$4\pi\mu = 2\pi nhc/e,$$

where  $n$  is the characteristic of the nodal line that ends there, or the sum of the characteristics of all nodal lines ending there when there is more than one. Thus at the end point there will be a magnetic pole of strength

$$\mu = \frac{1}{2}nhc/e.$$

Our theory thus allows isolated magnetic poles, but the strength of such poles must be quantised, the quantum  $\mu_0$  being connected with the electronic charge  $e$  by

$$hc/e\mu_0 = 2. \quad (9)$$

This equation is to be compared with (1). The theory also requires a quantisation of electric charge, since any charged particle moving in the field of a pole of strength  $\mu_0$  must have for its charge some integral multiple (positive or negative) of  $e$ , in order that wave functions describing the motion may exist.

#### *§ 4. Electron in Field of One-Quantum Pole.*

The wave functions discussed in the preceding section, having nodal lines ending on magnetic poles, are quite proper and amenable to analytic treatment by methods parallel to the usual ones of quantum mechanics. It will perhaps help the reader to realise this if a simple example is discussed more explicitly.

Let us consider the motion of an electron in the magnetic field of a one-quantum pole when there is no electric field present. We take polar co-ordinates  $\tau, \theta, \phi$ , with the magnetic pole as origin. Every wave function must now have a nodal line radiating out from the origin.

We express our wave function  $\psi$  in the form (3), where  $\beta$  is some non-integrable phase having derivatives  $\kappa$  that are connected with the known electromagnetic field by equations (6). It will not, however, be possible to obtain  $\kappa$ 's satisfying these equations all round the magnetic pole. There must be some singular line radiating out from the pole along which these equations are not satisfied, but this line may be chosen arbitrarily. We may choose it to be the same as the nodal line for the wave function under consideration, which would result in  $\psi_1$  being continuous. This choice, however, would mean different  $\kappa$ 's for different wave functions (the difference between any two being, of course, the four-dimensional gradient of a scalar, except on the singular lines). This would perhaps be inconvenient and is not really necessary. We may express all our wave functions in the form (3) with the same  $e^{i\beta}$ , and then those wave functions whose nodal lines do not coincide with the singular line for the  $\kappa$ 's will correspond to  $\psi_1$ 's having a certain kind of discontinuity on this singular line, namely, a discontinuity just cancelling with the discontinuity in  $e^{i\beta}$  here to give a continuous product.

The magnetic field  $\mathbf{H}$ , lies along the radial direction and is of magnitude  $\mu_0/\tau^2$ , which by (9) equals  $1/2hc/e\tau^2$ . Hence, from equations (7), curl  $\kappa$  is radial and of magnitude  $1/2\tau^2$ . It may now easily be verified that a solution of the whole of equations (7) is

$$\kappa_0 = 0, \quad \kappa_\tau = \kappa_\theta = 0, \quad \kappa_\phi = 1/2\tau \cdot \tan \frac{1}{2} \theta, \quad (10)$$

where  $\kappa_\tau, \kappa_\theta, \kappa_\phi$ , are the components of  $\kappa$  referred to the polar co-ordinates. This solution is valid at all points except along the line  $\theta = \pi$ , where  $\kappa_\phi$ , become infinite in such a way that  $\int(\kappa, d\mathbf{s})$  round a small curve encircling this line is  $2\pi$ . We may refer all our wave functions to this set of  $\kappa$ 's.

Let us consider a stationary state of the electron with energy  $W$ . Written non-relativistically, the wave equation is

$$-h^2/2m \cdot \nabla^2 \psi = W\psi.$$

If we apply the rule expressed by equation (5), we get as the wave equation for  $\psi_1$

$$-h^2/2m \cdot \left\{ \nabla^2 + i(\kappa, \nabla) + i(\nabla, \kappa) - \kappa^2 \right\} \psi_1 = W\psi_1. \quad (11)$$

The values (10) for the  $\kappa$ 's give

$$\begin{aligned} (\kappa, \nabla) &= (\nabla, \kappa) = \kappa_\phi \frac{1}{\tau \sin \theta} \frac{\partial}{\partial \phi} = \frac{1}{4\tau^2} \sec^2 \frac{1}{2} \theta \frac{\partial}{\partial \phi} \\ \kappa^2 &= \kappa_\phi^2 = \frac{1}{4\tau^2} \tan^2 \frac{1}{2} \theta, \end{aligned}$$

so that equation (11) becomes

$$-\frac{h^2}{2m} \left\{ \nabla^2 + \frac{i}{2\tau^2} \sec^2 \frac{1}{2} \theta \frac{\partial}{\partial \phi} - \frac{1}{4\tau^2} \tan^2 \frac{1}{2} \theta \right\} \psi_1 = W\psi_1.$$

We now suppose  $\psi_1$  to be of the form of a function  $f$  of  $\tau$  only multiplied by a function  $S$  of  $\theta$  and  $\phi$  only, i.e.,

$$\psi_1 = f(\tau)S(\theta\phi).$$

This requires

$$\left\{ \frac{d^2}{d\tau^2} + \frac{2}{\tau} \frac{d}{d\tau} - \frac{\lambda}{\tau^2} \right\} f = -\frac{2mW}{h^2} f, \quad (12)$$

$$\begin{aligned} \left\{ \frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \sin \theta \frac{\partial}{\partial \theta} + \frac{1}{\sin^2 \theta} \frac{\partial^2}{\partial \phi^2} + \frac{1}{2} i \sec^2 \frac{1}{2} \theta \frac{\partial}{\partial \phi} \right. \\ \left. - \frac{1}{4} \tan^2 \frac{1}{2} \theta \right\} S = -\lambda S, \end{aligned} \quad (13)$$

where  $\lambda$  is a number.

From equation (12) it is evident that there can be no stable states for which the electron is bound to the magnetic pole, because the operator on the left-hand side contains no constant with the dimensions of a length.

This result is what one would expect from analogy with the classical theory. Equation (13) determines the dependence of the wave function on angle. It may be considered as a generalisation of the ordinary equation for spherical harmonies.

The lowest eigenvalue of (13) is  $\lambda = 1/2$ , corresponding to which there are two independent wave functions

$$S_a = \cos \frac{1}{2} \theta, \quad S_b = \sin \frac{1}{2} \theta e^{i\phi},$$

as may easily be verified by direct substitution. The nodal line for  $S_a$  is  $\theta = \pi$ , that for  $S_b$ , is  $\theta = 0$ . It should be observed that  $S_a$  is continuous everywhere, while  $S_b$ , is discontinuous for  $\theta = \pi$ , its phase changing by  $2\pi$  when one goes round a small curve encircling the line  $\theta = \pi$ . This is just what is necessary in order that both  $S_a$  and  $S_b$ , when multiplied by the  $e^{i\beta}$  factor, may give continuous wave functions  $\psi$ . The two  $\psi$ 's that we get in this way are both on the same footing and the difference in behaviour of  $S_a$  and  $S_b$ , is due to our having chosen  $\kappa$ 's with a singularity at  $\theta = \pi$ .

The general eigenvalue of (13) is  $\lambda = n^2 + 2n + \frac{1}{2}$ . The general solution of this wave equation has been worked out by I. Tamm.<sup>9</sup>

### § 5. Conclusion.

Elementary classical theory allows us to formulate equations of motion for an electron in the field produced by an arbitrary distribution of electric charges and magnetic poles. If we wish to put the equations of motion in the Hamiltonian form, however, we have to introduce the electromagnetic potentials, and this is possible only when there are no isolated magnetic poles. Quantum mechanics, as it is usually established, is derived from the Hamiltonian form of the classical theory and therefore is applicable only when there are no isolated magnetic poles.

The object of the present paper is to show that quantum mechanics does not really preclude the existence of isolated magnetic poles. On the contrary, the present formalism of quantum mechanics, when developed naturally without the imposition of arbitrary restrictions, leads inevitably to wave equations whose only physical interpretation is the motion of an electron in the field of a single pole. This new development requires *no change whatever* in the formalism when expressed in terms of abstract symbols

---

<sup>9</sup>Appearing probably in 'Z. Physik.'

denoting states and observables, but is merely a generalisation of the possibilities of representation of these abstract symbols by wave functions and matrices. Under these circumstances one would be surprised if Nature had made no use of it.

The theory leads to a connection, namely, equation (9), between the quantum of magnetic pole and the electronic charge. It is rather disappointing to find this reciprocity between electricity and magnetism, instead of a purely electronic quantum condition, such as (1). However, there appears to be no possibility of modifying the theory, as it contains no arbitrary features, so presumably the explanation of (1) will require some entirely new idea.

The theoretical reciprocity between electricity and magnetism is perfect. Instead of discussing the motion of an electron in the field of a fixed magnetic pole, as we did in § 4, we could equally well consider the motion of a pole in the field of fixed charge. This would require the introduction of the electromagnetic potentials  $B$  satisfying

$$\mathbf{E} = \text{curl } \mathbf{B}, \quad \mathbf{H} = \frac{1}{c} \frac{\partial \mathbf{B}}{\partial t} + \text{grad } \mathbf{B}_0,$$

to be used instead of the  $A$ 's in equations (6). The theory would now run quite parallel and would lead to the same condition (9) connecting the smallest pole with the smallest charge.

There remains to be discussed the question of why isolated magnetic poles are not observed. The experimental result (1) shows that there must be some cause of dissimilarity between electricity and magnetism (possibly connected with the cause of dissimilarity between electrons and protons) as the result of which we have, not  $\mu_0 = e$ , but  $\mu_0 = 137/2 \cdot e$ . This means that the attractive force between two one-quantum poles of opposite sign is  $(137/2)^2 = 4692\frac{1}{4}$  times that between electron and proton. This very large force may perhaps account for why poles of opposite sign have never yet been separated.



A 1,500,000 Volt Electrostatic Generator

PROCEEDING of the AMERICAN PHYSICAL SOCIETY  
Minutes of the Schenectady Meeting,  
September 10, 11 and 12, 1931

Robert J. Van De Graaff  
Princeton University  
(Received 1931)

10. The application of extremely high potentials to discharge tubes affords a powerful means for the investigation of the atomic nucleus and other fundamental problems. The electrostatic generator here described was developed to supply suitable potentials for such investigations. In recent preliminary trials, spark-gap measurements showed a potential of approximately 1,500,000 volts, the only apparent limit being brush discharge from the whole surface of the 24-inch spherical electrodes. The generator has the basic advantage of supplying a direct steady potential, thus eliminating certain difficulties inherent in the application of non-steady high potentials. The machine is simple inexpensive, and portable. An ordinary lamp socket furnishes the only power needed. The apparatus is composed of two identical units, generating opposite potentials. The high potential electrode of each unit consists of a 24-inch hollow copper sphere mounted upon a 7 foot upright Pyrex rod. Each sphere is charged by a silk belt running between a pulley in its interior and a grounded motor driven pulley at the base of the

rod. The ascending surface of the belt is charged near the lower pulley by a brush discharge, maintained by 10.000 volt transformer kenotron set, and is subsequently discharged by points inside the sphere.

## The Existence of a Neutron

J. Chadwick

(Received 1932)

It was shown by Bothe and Becker that some light elements when bombarded by  $\alpha$ -particles of polonium emit radiations which appear to be of the  $\gamma$ -ray type. The element beryllium gave a particularly marked effect of this kind, and later observations by Bothe, by Mme. Curie-Joliot and by Webster showed that the radiation excited in beryllium possessed a penetrating power distinctly greater than that of any  $\gamma$ -radiation yet found from the radioactive elements. In Webster's experiments the intensity of the radiation was measured both by means of the Geiger-Muller tube counter and in a high pressure ionization chamber. He found that the beryllium radiation had an absorption coefficient in lead of about  $0.22 \text{ cm}^{-1}$  as measured under his experimental conditions. Making the necessary corrections for these conditions, and using the results of Gray and Tarrant to estimate the relative contributions of scattering, photoelectric absorption, and nuclear absorption in the absorption of such penetrating radiation, Webster concluded that the radiation had a quantum energy of about  $7 \times 10^6$  electron volts. Similarly he found that the radiation from boron bombarded by  $\alpha$ -particles of polonium consisted in part of a radiation rather more penetrating than that from beryllium, and he estimated the quantum energy of this component as about  $10 \times 10^6$  electron volts. These conclusions agree quite well with the supposition that the radiations arise by the capture of the  $\alpha$ -particle into the berillium (or boron) nucleus and emission of the surplus energy as a quantum of radiation.

The radiations showed, however, certain peculiarities, and at my request the beryllium radiation was passed into an expansion chamber and several photographs were taken. No unexpected phenomena were observed though, as will be seen later, similar experiments have now revealed some rather striking events. The failure of these early experiments was partly due to the weakness of the available source of polonium, and partly to the experimental arrangement, which as it now appears, was not very suitable.

Quite recently, Mme. Curie-Joliot and M. Joliot made the very striking observation that these radiation from berillium and from boron were able to eject protons with considerable velocities from matter containing hydrogen. In their experiments the radiation from beryllium was passed through a thin window into an ionisation vessel containing air at room pressure. When paraffin wax, or other matter containing hydrogen, was placed in front of the window, the ionisation in the vessel was increased, in some cases as much as doubled. The effect appeared to be due to the ejection of protons, and from further experiment they showed that the protons had ranges in air up to about 26 cm, corresponding to a velocity of nearly  $3 \times 10^9$  cm per second. They suggested that energy was transferred from the beryllium to the proton by a process similar to the Compton effect with electrons, and they estimated that the beryllium radiation had a quantum energy of about  $50 \times 10^6$  electron volts. The range of the protons ejected by the boron radiation was estimated to be about 8 cm in air, giving on a Compton process an energy of about  $35 \times 10^6$  electron volts for the effective quantum.<sup>1</sup>

There are two grave difficulties in such an explanation of this phenomenon. Firstly, it is now well established that the frequency of scattering of high energy quanta by electrons is given with fair accuracy by the Klein-Nishina formula, and this formula should also apply to the scattering of quanta by a proton. The observed frequency of the proton scattering is, however, many thousand times greater than that predicted by this formula. Secondly, it is difficult to account for the production of a quantum of  $50 \times 10^6$  electron volts from the interaction of a beryllium nucleus and an  $\alpha$ -particle of kinetic energy of  $5 \times 10^6$  electron volts. The process which will give the greatest amount of energy available for radiation is the capture of the  $\alpha$ -particle by the beryllium nucleus,  $\text{Be}^9$ , and its incorporation in the nuclear structure to form a carbon nucleus  $\text{C}^{13}$ . The mass defect of the  $\text{C}^{13}$  nucleus is known both from data supplied by measurements of the artificial disintegration of boron  $\text{B}^{10}$  and from observations of the band spectrum of

---

<sup>1</sup>Many of the arguments of the subsequent discussion apply equally to both radiations, and the term "beryllium radiation" may often be taken to include the boron radiation.

carbon; it is about  $10 \times 10^6$  electron volts. The mass defect of Be<sup>9</sup> is not known, but the assumption that it is zero will give a maximum value for the possible change of energy in the reaction  $\text{Be} + \alpha \rightarrow \text{C}^{13} + \text{quantum}$ . On this assumption it follows that the energy of the quantum emitted in such a reaction cannot be greater than about  $14 \times 10^6$  electron volts. It must, of course, be admitted that this argument from mass defects is based on the hypothesis that the nuclei are made as far as possible of  $\alpha$  particles; that the Be<sup>9</sup> nucleus constants of 2  $\alpha$ -particles + 1 proton + 1 electron and the C<sup>13</sup> nucleus of 3  $\alpha$ -particles + 1 proton + 1 electron. So far as the lighter nuclei are concerned, this assumption is supported by the evidence from experiments on artificial disintegration, but is no general proof.

Accordingly, I made further experiments to examine the properties of the radiation excited in beryllium. It was found that radiation ejects particles not only from hydrogen but from all other light elements which were examined. The experimental results were very difficult to explain on the hypothesis that the beryllium radiation was a quantum radiation, but followed immediately if it were supposed that the radiation consisted of particles of mass nearly equal to that of a proton and with no net charge, or neutron. .

..

#### OBSERVATION OF RECOIL ATOMS

The properties of the beryllium radiation were first examined by means of the valve counter used in the work on the artificial disintegration by  $\alpha$ -particles and described fully there. Briefly, it consists of a small ionisation chamber connected to a valve amplifier. The sudden production of ions in the chamber by the entry of an ionising particle is detected by means of an oscillograph connected in the output circuit of the amplifier. The deflections of the oscillograph were recorded photographically on a film of bromide paper.

The source of polonium was prepared from a solution of radium (D + E + F) by deposition on a disc of silver. The disc had a diameter of 1 cm and was placed close to a disc of pure beryllium of 2 cm diameter, and both were enclosed in a small vessel which could be evacuated

[Fig. 1]. The first ionisation chamber used had an opening of 13 mm covered with aluminium foil of 4.5 cm air equivalent, and a depth of 15

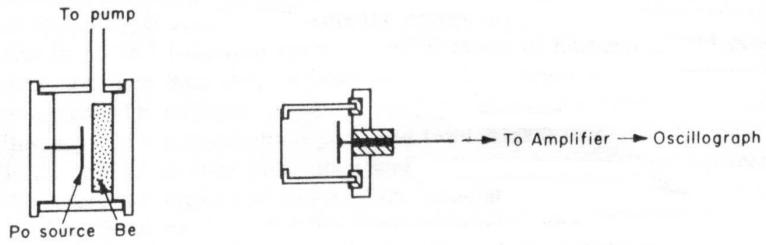


Figure 1:

mm. This chamber had a very low natural effect, giving on the average only about 7 deflections per hour.

When the source vessel was placed in front of the ionisation chamber, the number of deflections immediately increased. For a distance of 3 cm between the beryllium and the counter the number of deflections was nearly 4 per minute. Since the number of deflections remained sensibly the same when thick metal sheets, even as much as 2 cm of lead, were interposed between the source vessel and the counter, it was clear that these deflections were due to penetrating radiation emitted from the beryllium. It will be shown later that the deflections were due atoms of nitrogen set in motion by the impact of the beryllium radiation.

When a sheet of paraffin wax about 2 mm thick was interposed in the path of the radiation just in front of the counter, the number of deflections recorded by the oscillosograph increased markedly. This increase was due to particles ejected from the paraffin wax so as to pass into the counter.

By placing absorbing screens of aluminium between the wax and the counter the absorption curve shown in [Fig. 2], curve A, was obtained.

From this curve it appears that the particles have a maximum range of just over 40 cm of air, assuming that an Al foil of 1.64 mg. per square centimetre is equivalent to 1 cm of air. By comparing the sizes of the deflections (proportional to the number of ions produced in the chamber) due to these particles with those due to protons of about the same range it was obvious that the particles were protons. From the range-velocity curve for protons we deduce therefore that the maximum velocity imparted to a proton by the beryllium radiation is about  $3 \cdot 3 \times 10^9$  cm per second, corresponding to an energy of about  $5 \cdot 7 \times 10^6$  electron volts.

The effect of exposing other elements to the beryllium radiation was

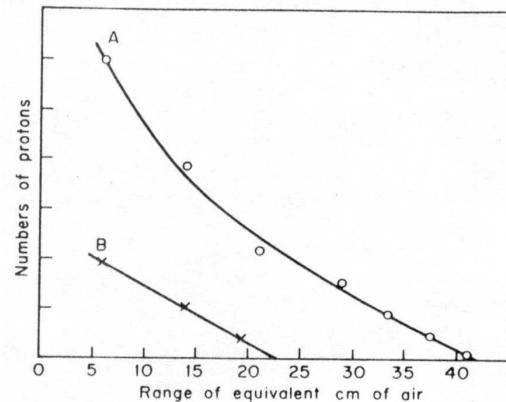


Figure 2:

then investigated. An ionisation chamber was used with an opening covered with a gold foil of 0.5 mm air equivalent. The element to be examined was fixed on a clean brass plate and placed very close to the counter opening. In this way, lithium, beryllium, boron, carbon and nitrogen, as paracyanogen, were tested. In each case the number of deflections observed in the counter increased when the element was bombarded by the beryllium radiation. The ranges of the particles ejected from these elements were quite short, of the order of some millimetres in air. The deflections produced by them were of different sizes, but many of them were large compared with the deflection produced even by a slow proton. The particles therefore have a large ionising power and are probably in each case recoil atoms of the elements. Gases were investigated by filling the ionisation chamber with the required gas by circulation for several minutes. Hydrogen, helium, nitrogen, oxygen, and argon were examined in this way. Again, in each case deflections were observed which were attributed to the production of recoil atoms in the different gases. For a given position of the beryllium source relative to the counter, the number of recoil atoms was roughly the same for each gas. This point will be referred to later. It appears then that the beryllium radiation can impart energy to the atoms of matter through which it passes and that the chance of an energy transfer does not vary widely from one element to another.

It has shown that protons are ejected from paraffin wax with energies up to a maximum of about  $5.7 \times 10^6$  electron volts. If the ejection be ascribed

to a Compton recoil from a quantum of radiation, then the energy of the quantum must be about  $55 \times 10^6$  electron volts, for the maximum energy which can be given to a mass  $m$  by a quantum  $h\nu$  is  $\frac{2}{2+mc^2/h\nu} \cdot h\nu$ . The energies of the recoil atoms produced by this radiation by the same process in other elements can be readily calculated. For example, the nitrogen recoil atoms should have energies up to a maximum of 450,000 electron volts. Taking the energy to form a pair of ions in air as 35 electron volts, the recoil atoms of nitrogen should produce not more than about 13,000 pairs of ions. Many of the deflections observed with nitrogen, however, corresponded to far more ions than this; some of the recoil atoms produced from 30,000 to 40,000 ion pairs. In the case of the other elements a similar discrepancy was noted between the observed energies and ranges of the recoil atoms and the values calculated on the assumption that the atoms were set in motion by recoil from a quantum of  $55 \times 10^6$  electron volts. The energies of the recoil atoms were estimated from the number of ions produced in the counter, as given by the size of the oscillograph deflections. A sufficiently good measurement of the ranges could be made either by varying the distance between the element and the counter or by interposing thin screens of gold between the element and the counter.

The nitrogen recoil atoms were also examined, in collaboration with Dr. N. Feather, by means of the expansion chamber. The source vessel was placed immediately above an expansion chamber of the Shimizu type, so that a large proportion of the beryllium radiation traversed the chamber. A large number of recoil tracks was observed in the course of a few hours. Their range, estimated by eye, was sometimes as much as 5 or 6 mm, in the chamber, or, correcting for the expansion, about 3 mm in standard air. These visual estimates were confirmed by a preliminary series of experiments by Dr. Feather with a large automatic expansion chamber, in which photographs of the recoil tracks in nitrogen were obtained. Now the ranges of recoil atoms of nitrogen of different velocities have been measured by Blackett and Lees. Using their results we find that the nitrogen recoil atoms produced by the beryllium radiation may have a velocity of at least  $4 \times 10^8$  cm per second, corresponding to an energy of about  $1 \cdot 2 \times 10^6$  electron volts. In order that the nitrogen nucleus should acquire such an energy in a collision with a quantum of radiation, it is necessary to assume that the energy of the quantum should be about  $90 \times 10^6$  electron volts, if energy and momentum are conserved in the collision. It has been shown that a quantum of  $55 \times 10^6$  electron volts is sufficient to explain the hydrogen collisions. In general, the experimental results show that if the recoil atoms are to be

explained by collision with a quantum, we must assume a larger and larger energy for the quantum as the mass of the struck atom increases.

### THE NEUTRON HYPOTHESIS

It is evident that we must either relinquish the application of the conservation of energy and momentum in collisions or adopt another hypothesis about the nature of the radiation. If we suppose that the radiation is not a quantum radiation, but consists of particles of mass very nearly equal to that of the proton, all the difficulties connected with the collisions disappear, both with regard to their frequency and to the energy transfer to different masses. In order to explain the great penetrating power of the radiation we must further assume that the particle has no net charge. We may suppose it to consist of a proton and an electron in close combination, the "neutron" discussed by Rutherford in his Bakerian Lecture of 1920.

When such neutrons pass through matter they suffer occasionally close collisions with the atomic nuclei and so give rise to the recoil atoms which are observed. Since the mass of the neutron is equal to that of the proton, the recoil atoms produced when the neutrons pass through matter containing hydrogen will have all velocities up to a maximum which is the same as the maximum velocity of the neutrons. The experiments showed that the maximum velocity of the protons ejected from paraffin wax was about  $3.3 \times 10^9$  cm per second. This is therefore the maximum velocity of the neutrons emitted from beryllium bombarded by  $\alpha$ -particles of polonium. From this we can now calculate the maximum energy which can be given by a colliding neutron to other atoms, and we find that the results are in fair agreement with the energies observed in the experiments. For example, a nitrogen atom will acquire in a head-on collision with the neutron of mass 1 and velocity  $3.3 \times 10^9$  cm per second a velocity of  $4.4 \times 10^8$  cm per second, corresponding to an energy of  $1.4 \times 1066$  electron volts, a range of about 3.3 mm in air, and a production of ions of about 40,000 pairs. Similarly, an argon atom may acquire an energy of  $0.54 \times 10^6$  electron volts, and produce about 15,000 ion pairs. Both these values are in good accordance with experiment. It is possible to prove that the mass of the neutron is roughly equal to that of the proton, by combining the evidence from the hydrogen collisions with that of the nitrogen collisions. In the succeeding paper, Feather records experiments in which about 100 tracks of nitrogen recoil atoms have been photographed

in the expansion chamber. The measurement of the tracks shown that the maximum range of the recoil atoms is 3 · 5 mm in air 15° C and 760 mm pressure, corresponding to a velocity of  $4 \cdot 7 \times 10^8$  cm per second according to Blackett and Lees. If  $M$ ,  $V$  be the mass and velocity of the neutron then the maximum velocity given to a hydrogen atom is

$$u_p = \frac{2M}{M+1} \cdot V,$$

and the maximum velocity given to a nitrogen atom is

$$u_n = \frac{2M}{M+14} \cdot V,$$

whence

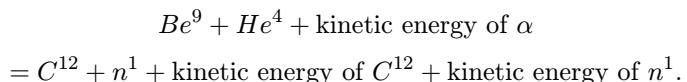
$$\frac{M+14}{M+1} = \frac{u_p}{u_n} = \frac{3 \cdot 3 \times 10^9}{4 \cdot 7 \times 10^8},$$

and

$$M = 1 \cdot 15.$$

The total error in the estimation of the velocity of the nitrogen recoil atom may easily be about 10 per cent., and it is legitimate to conclude that the mass of the neutron is very nearly the same as the mass of the proton.

We have now to consider the production of the neutrons from beryllium by the bombardment of the  $\alpha$ -particles. We must suppose that an  $\alpha$ -particle is captured by a  $\text{Be}^9$  nucleus with the formation of a carbon  $\text{C}^{12}$  nucleus and the emission of a neutron. The process is analogous to the well-known artificial disintegrations., but a neutron is emitted instead of a proton. The energy relations of this process cannot be exactly deduced, for the masses of the  $\text{Be}^9$  nucleus and the neutron are not known accurately. It is, however, easy to show that such a process fits the experimental facts. We have



If we assume that the beryllium nucleus consists of two  $\alpha$ -particles and a neutron, then its mass cannot be greater than the sum of the masses of these particles, for the binding energy corresponds to a defect of mass.

The energy equation becomes

$$\begin{aligned} & (8 \cdot 00212 + n^1) + 4 \cdot 00106 + \text{K.E. of } \alpha > 12 \cdot 0003 + n^1 \\ & + \text{K.E. of } \text{C}^{12} + \text{K.E. of } N^1 \end{aligned}$$

or

$$K.E. \text{ of } n^1 < K.E. \text{ of } \alpha + 0 \cdot 003 - K.E. \text{ of } C^{12}.$$

Since the kinetic energy of the  $\alpha$ -particle of polonium is  $5 \cdot 25 \times 10^6$  electron volts, it follows that the energy of emission of the neutron cannot be greater than about  $8 \times 10^6$  electron volts. The velocity of the neutron must therefore be less than  $3 \cdot 9 \times 10^9$  cm per second. We have seen that the actual maximum velocity of the neutron is about  $3 \cdot 3 \times 10^9$  cm per second, so that the proposed disintegration process is compatible with observation.

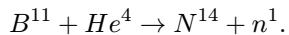
A further test of the neutron hypothesis was obtained by examining the radiation emitted from beryllium in the opposite direction to the bombarding  $\alpha$ -particles. The source vessel [Fig. 1] was reversed so that a sheet of paraffin wax in front of the counter was exposed to the "backward" radiation from the beryllium. The maximum range of the protons ejected from the wax was determined as before, by counting the numbers of protons observed through different thicknesses of aluminium interposed between the wax and the counter. The absorption curve obtained is shown in curve B, [Fig. 74-2]. The maximum of the protons was about 22 cm in air, corresponding to a velocity of about  $2 \cdot 74 \times 10^9$  cm per second. Since the polonium source was only about 2 mm away from the beryllium, this velocity should be compared with that of the neutrons emitted not at 180 degrees but at an angle not much greater than 90° to the direction of the incident  $\alpha$ -particles. A simple calculation shows that the velocity of the neutron emitted at 90° when an  $\alpha$ -particle of full range is captured by a beryllium nucleus should be  $2 \cdot 77 \times 10^9$  cm per second, taking the velocity of the neutron emitted at 0 degree in the same process as  $3 \cdot 3 \times 10^9$  cm per second. The velocity found in the above experiment should be less than this, for the angle of emission is slightly greater than 90 degrees. The agreement with calculation is as good as can be expected from such measurements.

## THE NATURE OF THE NEUTRON

It has been shown that the origin of the radiation from beryllium bombarded by  $\alpha$ -particles and the behaviour of the radiation, so far as its interaction with atomic nuclei is concerned, receive a simple explanation on the assumption that the radiation consists of particles of mass nearly equal to that of the proton which have no charge. The simplest hypothesis one can make about the nature of the particle is to suppose that it consists of

a proton and an electron in close combination, giving a net charge 0 and a mass which should be slightly less than the mass of the hydrogen atom. This hypothesis is supposed by an examination of the evidence which can be obtained about the mass of the neutron.

As we have seen, a rough estimate of the mass of the neutron was obtained from measurements of its collisions with hydrogen and nitrogen atoms, but such measurements cannot be made with sufficient accuracy for the present purpose. We must turn to a consideration of the energy relations in a process in which a neutron is liberated from an atomic nucleus; if the masses of the atomic nuclei concerned in the process are accurately known, a good estimate of the mass of the neutron can be deduced. The mass of the beryllium nucleus has, however, not yet been measured, and, as was shown [earlier], only general conclusions can be drawn from this reaction. Fortunately, there remains the case of boron. It was stated in [the first section] that boron bombarded by  $\alpha$ -particles of polonium also emits a radiation which ejects protons from materials containing hydrogen. Further examination showed that this radiation behaves in all respects like that from beryllium, and it must therefore be assumed to consist of neutrons. It is probable that the neutrons are emitted from the isotope  $B^{11}$ , for we know that the isotope  $B^{10}$  disintegrates with the emission of a proton. The process of disintegration will then be



The masses of  $B^{11}$  and  $N^{14}$  are known from Aston's measurements, and the further data required for the deduction of the mass of the neutron can be obtained by experiment.

In the source vessel of [Fig. 1] the beryllium was replaced by a target of powdered boron, deposited on a graphite plate. The range of the protons ejected by the boron radiation was measured in the same way as with the beryllium radiation. The effects observed were much smaller than with beryllium, and it was difficult to measure the range of the protons accurately. The maximum range was about 16 cm in air, corresponding to a velocity of  $2.5 \times 10^9$  cm per second. This then is the maximum velocity of the neutron liberated from boron by an  $\alpha$ -particle of polonium of velocity  $1.59 \times 10^9$  cm per second assuming that momentum is conserved in the collision, the velocity of the recoiling  $N^{14}$  nucleus can be calculated, and we then know the kinetic energies of all particles concerned in the disintegration process. The energy equation of the process is

$$\text{Mass of } B^{11} + \text{mass of } He^4 + K.E. \text{ of } He^4$$

$$= \text{mass of } N^{14} + \text{mass of } n^1 + K.E. \text{ of } N^{14} K.E. \text{ of } n^1.$$

The masses are  $B^{14} = 11.00825 \pm 0.0016$ ;  $He^4 = 4.00106 \pm 0.0006$ ;  $N^{14} = 14.0042 \pm 0.0028$ . The kinetic energies in mass units are  $\alpha$ -particle =  $0.00565$ ; neutron =  $0.0035$ ; and nitrogen nucleus =  $0.00061$ . We find therefore that the mass of the neutron is  $1.0067$ . The errors quoted for the mass measurements are those given by Aston. They are the maximum errors which can be allowed in his measurements, and the probable error may be taken as about one-quarter of these. Allowing for the errors in the mass measurements it appears that the mass of the neutron cannot be less than  $1.003$ , and that it probably lies between  $1.005$  and  $1.008$ .

Such a value for the mass of the neutron is to be expected if the neutron consists of a proton and an electron, and it lends strong support to this view. Since the sum of the masses of the proton and electron is  $1.0078$ , the binding energy, or mass defect, of the neutron is about 1 to 2 million electron volts. This is quite a reasonable value. We may suppose that the proton and electron form a small dipole, or we may take the more attractive picture of a proton embedded in an electron. On either view, we may expect the "radius" of the neutron to be a few times  $10^{-13}$  cm

### THE PASSAGE OF THE NEUTRON THROUGH MATTER

The electrical field of a neutron of this kind will clearly be extremely small except at very small distances of the order of  $10^{-12}$  cm. In its passage through matter the neutron will be deflected unless it suffers an intimate collision with a nucleus. The potential of a neutron in the field of a nucleus may be represented roughly by [Fig. 3]. The radius of the collision area for sensible deflection of the neutron will be little greater than the radius of the nucleus. Further, the neutron should be able to penetrate the nucleus easily, and it may be that the scattering of the neutrons will be largely due to the internal field of the nucleus, or, in other words, that the scattered neutrons are mainly those which have penetrated the potential barrier. On these views we should expect the collision of a neutron with a nucleus to occur very seldom, and that the scattering will be roughly equal in all directions, at least as compared with the Coulomb scattering of a charged particle.

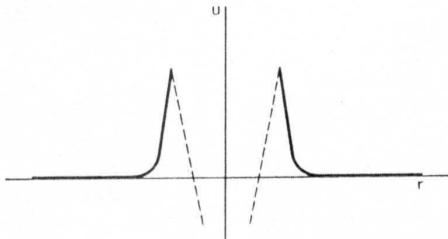


Figure 3:

These conclusions were confirmed in the following way. The source vessel, with Be target, was placed rather more than 1 inch from the face of a closed counter filled with air. [Fig. 1]. The number of deflections, or the number of nitrogen recoil atoms produced in the chamber, was observed for a certain time. The number observed was 190 per hour, after allowing for the natural effect. A block of lead 1 inch thick was then introduced between the source vessel and the counter. The number of deflections fell to 166 per hour. Since the number of recoil atoms produced must be proportional to the number of neutrons passing through the counter, these observations show that 13 per cent. of the neutrons had been absorbed or scattered in passing through 1 inch of lead.

Suppose that a neutron which passes within a distance  $p$  from the centre of the lead nucleus is scattered and removed from the beam. Then the fraction removed from the beam in passing through a thickness  $t$  of lead will be  $\pi p^2 n t$ , where  $n$  is the number of lead atoms per unit volume. Hence  $\pi p^2 n t = 0 \cdot 13$ , and  $p = 7 \times 10^{-13}$  cm. This value for the collision radius with lead seems perhaps rather small, but it is not unreasonable. We may compare it with the radii of the radioactive nuclei calculated from the disintegration constants by Gamow and Houtermans, viz., about  $7 \times 10^{-13}$  cm.

Similar experiments were made in which the neutron radiation was passed through blocks of brass and carbon. The values of  $p$  deduced in the same way were  $6 \times 10^{-13}$  cm and  $3 \cdot 5 \times 10^{-13}$  cm respectively.

The target areas for collision for some light elements were compared by another method. The second ionization chamber was used, which could be filled with different gases by circulation. The position of the source vessel was kept fixed relative to the counter, and the number of deflections was observed when the counter was filled in turn with hydrogen, nitrogen, oxy-

gen, and argon. Since the number of neutrons passing through the counter was the same in each case, the number of deflections should be proportional to the target area for collision, neglecting the effect of the material of the counter, and allowing for the fact that argon is monatomic. It was found that nitrogen, oxygen, and argon give about the same number of deflections; the target areas of nitrogen and oxygen are thus roughly equal, and the target area of argon is nearly twice that of these. With hydrogen the measurements were very difficult, for many of the deflections were very small owing to the low ionising power of the proton and the low density of the gas. It seems probable from the results that the target area of hydrogen is about two-thirds that of nitrogen or oxygen, but it may be rather greater than this.

There is yet little information about the angular distribution of the scattered neutrons. In some experiments kindly made for me by Dr.Gray and Mr.Lea, the scattering by lead was compared in the backward and forward directions, using the ionisation in a high pressure chamber to measure the neutrons. They found that the amount of scattering was about that to be expected from the measurements quoted above, and that the intensity per unit solid angle was about the same between  $30^\circ$  to  $90^\circ$  in the forward direction as between  $90^\circ$  to  $150^\circ$  in the backward direction. The scattering by lead is therefore not markedly anisotropic.

Two types of collision may prove to be of peculiar interest, the collision of a neutron with a proton and the collision with an electron. A detailed study of these collisions with an elementary particle is of special interest, for it should provide information about the structure and field of the neutron, whereas the other collisions will depend mainly on the structure of the atomic nuclei. Some preliminary experiments by Mr.Lea, using the pressure chamber to measure the scattering of neutrons by paraffin wax and by liquid hydrogen, suggest that the collision with a proton is more frequent than with other light atoms. This is not in accord with the experiments described above, but the results are at present indecisive. These collisions can be more directly investigated by means of the expansion chamber or by counting methods, and it is hoped to do so shortly.

The collision of a neutron with an electron has been examined in two ways, by the expansion chamber and by the counter. An account of the expansion chamber experiments is given by Mr.Dee in the third paper of this series. Mr.Dee has looked for the general ionisation produced by a large number of neutrons in passing through the expansion chamber, and also for the short electron tracks which should be the result of a very close collision between a neutron and electron. His results show that collisions

with electron are extremely rare compared even with those with nitrogen nuclei, and he estimates that a neutron can produce on the average not more than 1 ion pair in passing through 3 metres of air.

In the counter experiments a beam of neutrons was passed through a block of brass, 1 inch thick, and the maximum range of the protons ejected from paraffin wax by the emergent beam was measured. From this range the maximum velocity of the neutrons after travelling through the brass is obtained and it can be compared with the maximum velocity in the incident beam. No change in the velocity of the neutrons due to their passage through the brass could be detected. The accuracy of the experiment is not high, for the estimation of the end of the range of the protons was rather difficult. The results show that the loss of energy of a neutron in passing through 1 inch of brass is not more than about  $0 \cdot 4 \times 10^6$  electron volts. A path of 1 inch in brass corresponds as regards electron collisions to a path of nearly  $2 \times 10^4$  cm of air, so that result would suggest that a neutron loses less than 20 volts per centimetre path in air in electron collisions. This experiment thus lends general support to those with the expansion chamber, though it is of far inferior accuracy. we conclude that the transfer of energy from the neutron to electrons is of very rare occurrence. This is not unexpected. Bohr has shown on quite general ideas that collisions of a neutron with an electron should be very few compared with nuclear collisions. Massey, on plausible assumptions about the field of the neutron, has made a detailed calculation of the loss of energy to electrons, and finds also that it should be small, not more than 1 ion pair metre in air.

#### GENERAL REMARKS

It is interest to examine whether other elements, besides beryllium and boron, emit neutrons when bombarded by  $\alpha$ -particles. So far as experiments have been made, no case comparable with these two has been found. Some evidence was obtained of the emission of neutrons from fluorine and magnesium, but the effects were very small, rather less than 1 per cent. of the effect obtained from beryllium under the same conditions. there is also the possibility that some elements may emit neutrons spontaneously, e.g., potassium, which is known to emit a nuclear  $\beta$ -radiation accompanied by a more penetrating radiation. Again no evidence was found of the presence of neutrons, and it seems fairly certain that the penetrating type is, as has been assumed, a  $\gamma$ -radiation.

Although there is certain evidence for the emission of neutrons only in two cases of nuclear transformations, we must nevertheless suppose that the neutron is a common constituent of atomic nuclei. We may then proceed to build up nuclei out of  $\alpha$ -particles, and protons, and we are able to avoid the presence of uncombined electrons in a nucleus. This has certain advantages for, as is well known, the electrons in a nucleus have lost some of the properties which they have outside, e.g., their spin and magnetic moment. If the  $\alpha$ -particle, the neutron, and the proton are the only units of nuclear structure, we can proceed to calculate the mass defect or binding energy of a nucleus as the difference between the mass of the nucleus and the sum of the masses of the constituent particles. It is, however, by no means certain that the  $\alpha$ -particle and the neutron are the only complex particles in the nuclear structure, and therefore the mass defects calculated in this way may be the true binding energies of the nuclei. In this connection it may be noted that the examples of disintegration discussed by Dr. Feather in the next paper are not all of one type, and he suggests that in some cases a particle of mass 2 and charge 1, the hydrogen isotope recently reported by Urey, Brickwedde and Murphy, may be emitted. It is indeed possible that this particle also occurs as a unit of nuclear structure.

It has so far been assumed that the neutron is a complex particle consisting of a proton and an electron. This is the simplest assumption and it is supported by the evidence that the mass of the neutron is about 1.006, just a little less than sum of the masses of a proton and an electron. Such a neutron would appear to be the first step in the combination of the elementary particles towards the formation of a nucleus. It is obvious that this neutron may help us to visualise the building up of more complex structures, but the discussion of these matters will not be pursued further for such speculations, though not idle, are not at the moment very fruitful. It is, of course, possible to suppose that the neutron may be an elementary particle. This view has little to recommend it at present, except the possibility of explaining the statistics of such nuclei as  $N^{14}$ .

There remains to discuss the transformations which take place when an  $\alpha$ -particle is captured by a beryllium nucleus,  $Be^9$ . The evidence given here indicates that the main type of transformation is the formation of a  $C^{12}$  nucleus and the emission of a neutron. The experiments of Curie-Joliot and Joliot, of Auger, and of Dee show quite definitely that there is some radiation emitted by beryllium which is able to eject fast electrons in passing through matter. I have made experiments using the Geiger point counter to investigate this radiation and the results suggest that the electrons are produced by a  $\gamma$ -radiation. There are two distinct processes which may

give rise to such a radiation. In the first place, we may suppose that the transformation of  $\text{Be}^9$  to  $\text{C}^{12}$  takes place sometimes with the formation of an excited  $\text{C}^{12}$  nucleus which goes to the ground state with the emission of  $\gamma$ -radiation. This is similar to the transformations which are supposed to occur in some cases of disintegration with proton emission, e.g.,  $\text{B}^{10}4$ ,  $\text{F}^{19}$ ,  $\text{Al}^{27}$ , the majority of transformations occur with the formation of an excited nucleus, only in about one-quarter is the final state of the residual nucleus reached in one step. We should then have two groups of neutrons of different energies and a  $\gamma$ -radiation of quantum energy equal to the difference in energy of the neutron groups. The quantum energy of this radiation must be less than maximum energy of the neutrons emitted, about  $5 \cdot 7 \times 10^6$  electron volts. In the second place, we may suppose that occasionally the beryllium nucleus changes to a  $\text{C}^{13}$  nucleus and that the surplus energy is emitted as radiation. In this case the quantum energy of the radiation may be about  $10 \times 10^6$  electron volts.

It is of interest to note that Webster has observed a soft radiation from beryllium bombarded by polonium  $\alpha$ -particles, of energy about  $5 \times 10^5$  electron volts. This radiation may well be ascribed to the first of the two processes just discussed, and its intensity is of the right order. On the other hand, some of the electrons observed by Curie-Joliot and Joliot had energies of the order of 2 to  $10 \times 10^6$  volts, and Auger recorded one example of an electron of energy about  $6 \cdot 5 \times 10^6$  volts. These electrons may be due to a hard  $\gamma$ -radiation produced by the second type of transformation.<sup>2</sup>

It may be remarked that no electrons of greater energy than the above appear to be present. This is confirmed by an experiment made in this laboratory by Dr. Occhialini. Two tube counters were placed in a horizontal plane and the number of coincidences recorded by them was observed by means of the method devised by Rossi. The beryllium source was then brought up the plane of the counters so that the radiation passed through both counters in turn. No increase in the number of coincidences could be detected. It follows that there are few, if any,  $\beta$ -rays produced with energies sufficient to pass through the walls of both counters, a total of 4 mm brass; that is, with energies greater than about  $6 \times 10^8$  volts. This experiment further shows that the neutrons very rarely produce coincidences in tube counters under the usual conditions of experiment.

---

<sup>2</sup>Although the presence of fast electrons can be easily explained in this way, the possibility that some may be due to secondary effects of the neutrons must be lost sight of.

## The Electrostatic Production of High Voltage for Nuclear Investigations

R. J. Van de Graaf<sup>1</sup>, K. T. Compton and L. C. Van Atta,  
Massachusetts Institute of Technology  
(Received December 20, 1932)

### Abstract

The developments in nuclear physics emphasize the need of a new technique adapted to deliver enormous energies in concentrated form in order to penetrate or disrupt atomic nuclei. This may be achieved by a generator of current at very high voltage. Economy, freedom from the inherent defects of an impulsive, alternating or rippling source and the logic of simplicity point to an electrostatic generator as a suitable tool for this technique. Any such generator needs a conducting terminal, its insulating support and a means for conveying electricity to the terminal. These needs are naturally met by a hollow metal sphere supported on an insulator and charged by a belt conveying electricity from earth potential and depositing it within the interior of the sphere. Four models of such a generator are described, three being successive developments of generators operating in air, and designed respectively for 80,000, 1,500,000 and 10,000,000 volts, and the fourth being an essentially similar generator operating in a highly evacuated tank. Methods are described for depositing electric charge on the belts either by external or by self-excitation. The upper limit to the attainable voltage is set by the breakdown strength of the insulating medium surrounding the sphere, and by its size. The upper limit to the current is set by the rate at which belt area enters the sphere, carrying a surface density of charge whose upper limit is that which causes a breakdown field in the

---

<sup>1</sup>National Research Fellow at Princeton from September 1929 to September 1931.

surrounding medium, e.g., 30,000 volts per cm if the medium is air at atmospheric pressure. The voltage and the current each vary as the breakdown strength of the surrounding medium and the power output as its square. Also the voltage, current and power vary respectively as the 1st, 2nd and 3rd powers of the linear dimensions.

## 1 Introduction

Any basic development in science calls for a new technique fundamentally adapted to the new purpose. Such a basic development has been in progress for the last thirty-six years and it has now reached such a stage as to be recognized as a major field of physics,—nuclear physics. To justify this statement it is only necessary to point out that in the atomic nuclei reside all the positive electricity, much of the negative electricity and by far the greater part of the mass and energy in the universe. As yet we know relatively little about this world of electricity, mass and energy, and our attempts to produce an impression on it have been still less successful. Nevertheless, recent years have demonstrated the possibility of successful attack upon the nucleus by one weapon, the high-speed particle. The new technique which is therefore demanded is a powerful, controllable source of high-speed particles. In this paper is described a high-voltage generator which, with accessory discharge tubes and sources of ions, appears adapted to this technique.

Since progress in this new field of nuclear physics will require a great investment in apparatus, time and effort, it is important to consider carefully the possible techniques and whether this investment may be justified by the importance of the results to be expected. This leads to a consideration of the effects on the progress of physical theory of basic experimental developments initiated at the close of the last century.

About 1895 the discovery of *x*-rays and other experimental developments made possible an effective entry into the outer structure of the atom and the investigation of this aspect of atomic structure has been the main business of physics since that time. The distinguishing feature of those experimental innovations which led to the present knowledge of the outer atom is that they made possible physical observations on a scale of dimensions thousands of times smaller, with resultant energy concentrations thousands of times larger than were previously possible. Thus it became possible to deal individually with the basic physical entities, thereby removing the limitations of the statistical approach formerly necessary. Such observations showed clearly the insufficiency of classical laws and paved the way for quantum theory and for relativity.

Just as *x*-rays and other radiation phenomena of extranuclear electrons unfolded for us the extranuclear structure of atoms, so the discovery of radioactivity in 1896 opened the way to the first knowledge of the inner structure of the atomic nucleus. Furthermore, the high-speed particles from radioactive substances have themselves been the agencies through which was obtained the first evidence of the possibility of atomic transformation. Simultaneously with these experimental developments, the theory of relativity, through Einstein's principle of interconvertibility of mass and energy, has given a partial basis for guidance in interpretation and investigation of nuclear changes. These developments in nuclear physics have suggested with increasing emphasis the tremendous possibilities which should accrue from further development in the technique of nuclear investigation, through the use of swift particles of controllable nature and speed produced and applied through the agency of high voltage. In fact, the recent brilliant experiments of Cockcroft and Walton<sup>2</sup> have given concrete evidence that these expectations were justifiable. It may be hoped that experimental entry into the nucleus will extend the technique of physical exploration in a manner analogous to the extension which opened up the outer structure of the atom. It will again be possible to study a system thousands of times smaller, with energy concentrations correspondingly greater, further facilitating the study of individual rather than statistical processes. This may result in a system of nuclear mechanics accompanied by the same sort of broadening of basic scientific and philosophical ideas as accompanied the creation of quantum mechanics.

Within the almost unknown nuclear world of electricity and energy may also lie the explanation of certain extranuclear phenomena, which, modern quantum theory notwithstanding, still remain obscure. Just as the discovery of nuclear charge led immediately to the interpretation and simplification of that complicated group of chemical phenomena partially classified by the periodic table, so similarly it is possible that further discoveries regarding the atomic nucleus may lead to a similar interpretation and simplification of some extranuclear problems still outstanding.

## 2 Production of High Voltage

In approaching the problem of a high-voltage technique fundamentally adapted to meet the new demands in the most perfect and ultimate manner,

---

<sup>2</sup>Cockcroft and Walton, Nature **129**, 649 (April, 1932); Proc. Roy. Soc. **A137**, 229 (1932).

it is well to review the development of the high-voltage art to its present state<sup>3</sup>. Before the time of Faraday electrostatic generators were employed. However, industrial developments of the past hundred years have found their most suitable embodiment in applications of Faraday's principles of electromagnetism. Thus modern high-voltage technique has evolved almost completely under this influence. Step-up transformers have been used, with the addition of rectifiers and condensers when an approximately steady direct current was desired. These electromagnetic devices are admirably suited for the production of large currents within the general range of voltages corresponding to extranuclear phenomena. There are, however, inherent difficulties in the extension of such devices into the range of extremely high voltages which are demanded by nuclear physics. Such difficulties include the tremendous expense and size of such generators, necessitated by insulation requirements. There is also the limitation that the efficiency of high-voltage a.c. devices decreases rapidly as higher voltages are sought, because of the parasitic charging currents which are required to bring the apparatus to high potential at every cycle, even though no power is being drawn by it. The importance of this feature is not generally realized, but may be illustrated by the statement that the most favorable arrangement of the two terminals alone, neglecting the circuits, would require about 10,000 kva to impress 10,000,000 volts at 60 cycles, exclusive of useful output or corona leakage.

There are, in addition, numerous other important advantages of steady direct current over any current of a surging, alternating or rippling character. These include:

- (1) Possibility of obtaining strictly homogeneous beams of electrified particles.
- (2) Possibility of accurate focussing which becomes relatively more important as the voltage and therefore the length of discharge tubes is increased.
- (3) Elimination of stray radiation which arises in a variety of ways from a discharge tube and generating apparatus operating at unsteady voltages, and which renders difficult the careful shielding necessary to permit the use of the delicate instruments and sensitive amplifiers required in so many applications of these voltages.
- (4) Ability to use the ion source to full capacity since the useful voltage

---

<sup>3</sup>Lack of space prevents discussion of methods now used for generating high-energy radiations, but reference may be made to such well-known work as that of Coolidge, Tuve, Lauritsen, Cockcroft and Walton, Brasch and Lange, and Lawrence, and their collaborators.

is applied all of the time. (The order of advantage in this respect runs from the impulse generator and Tesla coil, which utilize the source only during roughly a millionth of the time, through the induction coil, the a.c. transformer, the transformer with rectifier, to finally, the electrostatic d.c. generator.)

(5) Ability to measure the high potentials accurately as, for instance, by the use of null or compensation methods.

(6) Elimination of breakdown in vacuum tubes due to reversal of the voltage,—a phenomenon inherently associated with the walls of vacuum tubes even under the best available conditions of evacuation.

(7) Ability to utilize the advantages of geometrical dissymmetry between positive and negative electrodes in vacuum in such a way that the field is minimum at the surface of the negative electrode, where difficulties from field currents are most serious,—an advantage which finds its maximum embodiment when the cathode is a hollow sphere surrounding a central spherical anode.

(8) In a variety of other ways through the elimination of time variations in the electrical conditions of the apparatus.

In view of these considerations it seemed desirable to develop an electrostatic high-voltage generator, since electrostatic methods yield directly a steady unidirectional voltage such as is desired. Maximum simplicity was sought in the design. The simplest terminal assembly appeared to be a sphere mounted on an insulating column. Since the sphere must be charged and since the process should be continuous the charge carrier should approach the sphere, enter it, and, after depositing its charge inside should return parallel to its path of approach. This immediately suggested the action of a belt, a device long used for the transmission of mechanical power.

The logic of the situation therefore pointed directly to a generator consisting of a hollow spherical conducting terminal supported on an insulating column, a moving belt to carry electric charge to the sphere, a device for depositing the charge onto the belt in a region of low potential remote from the sphere, and a device for removing this charge from the belt inside the sphere and transferring it to the sphere. A refinement of these essentials was the addition of an induction device whereby charge of the opposite sign was carried by the belt on its return journey, thus doubling the current output. A second refinement consisted of a self-exciting charging device whereby the entire generator could be made to operate independently of any external source of electricity. Not only does this device attain the desired result in what appears to be the simplest possible manner, but it is also interesting to note that the energy transformations in its operations are exceedingly

simple, consisting only in the transformation of the energy required to drive the belt into work done in separating and transferring electric charge from earth potential to sphere potential.

#### Historical note

The basic idea of a belt type of generator probably dates from Kelvin's famous water dropper<sup>4</sup>, and in fact, Kelvin suggested such a generator, in which charges would be carried to the electrode on a belt conveyer consisting of alternately insulated metal segments. Righi<sup>5</sup> made such a generator with the segmented belt carried through the sphere. Later Burboa<sup>6</sup> designed a generator with a belt functioning somewhat as an elongated disk of a Wimshurst machine, with a complicated set of inductors and brushes. Mention also should be made of a generator designed by Swann<sup>7</sup>, in which charge was conveyed by a succession of falling metal spheres, thus coming closer to Kelvin's original water dropper but with the added suggestion that this apparatus could be made to operate in vacua theoretically up to such voltages as would prevent the falling of the spheres by electrostatic forces. Still more recently Vollrath<sup>8</sup> constructed a similar generator in which the current is carried by an air blast of electrified dust in an insulating tube.

### 3 Principles of Operation

The simplest embodiment of these principles is illustrated schematically in Fig. 1, which is appropriate to a generator, where  $P$  and  $N$  are the positive and negative spherical terminals and the belts of silk, or any other flexible insulating material, are shown transporting positive and negative electricity, respectively, from the charging outfit connected with ground to the two spheres. The charge is "sprayed" onto each belt as it passes between a metallic surface and one or more sharp points so adjusted as to maintain a brush discharge from the points toward the surface. When connected as shown, one point sprays positive and the other sprays negative electricity onto its adjacent belt. Within each sphere the charge is drawn off the belt by adjacent sharp points and transferred to the spheres. Since the interior of the sphere is similar to the interior of a Faraday "ice pail" the charge

---

<sup>4</sup>John Gray, *Electrical Influence Machines*, Whittaker, London, 1890.

<sup>5</sup>John Gray, *Electrical Influence Machines*, Whittaker, London, 1890.

<sup>6</sup>Burboa, U. S. Patent Nos. 776, 997 (1904).

<sup>7</sup>Swann, J. Frank. Inst. **205**, 820 (1928).

<sup>8</sup>Vollrath, Phys. Rev. 42, 298 (1932).

passes readily between the charged belt and the sphere, irrespective of the potential of the sphere.

The voltage which is attainable in this device is limited only by the corona breakdown at the surface of the spheres, which depends in a known manner upon their size and varies somewhat with the degree of smoothness of their surfaces. The current output is equal to the rate at which charge is carried into the spheres and is therefore equal to the product of the surface

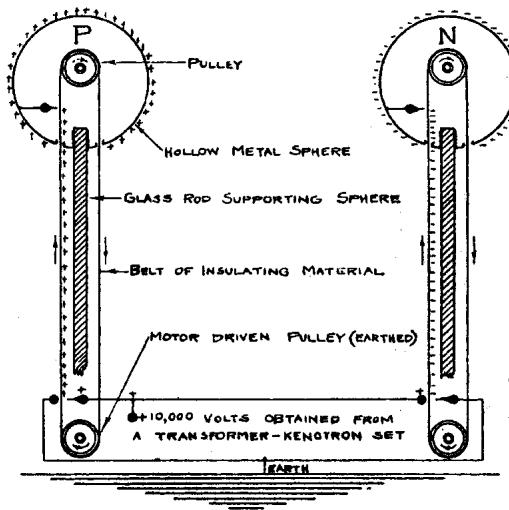


Figure 1:

density of charge on the belts multiplied by the areal velocity with which the belts enter the spheres. The upper limit for the surface density of charge on the belts is that which gives rise to an electric field equal to the "breakdown" strength of the surrounding air.

It is evident that other insulating media than air at atmospheric pressure might be used to advantage in the operation of such a generator<sup>9</sup>. We believe that its most useful embodiments will prove to be with operation in a high vacuum tank. In any case, the voltage is limited by the electrical breakdown of the surrounding medium whatever it may be, and the current output is determined by the width and velocity of the belt together with the charge density which can be placed on it.

---

<sup>9</sup>Barton, Mueller and Van Atta, *A Compact High-Potential Electrostatic Generator*, Phys. Rev. **42**, 901 A (1932).

In view of these considerations it is evident that the maximum voltage and also the current of such a generator each vary directly as the breakdown strength of the surrounding medium, so that the power output varies as the square of this breakdown strength. Since generally it will be desirable to provide as large a current output as possible, the belts will be designed to operate at the greatest practicable speed, and multiple belts will be placed as closely together as convenient. It is therefore evident that in any given insulating medium the voltage will vary as the first power, the current as the second power, and the power output as the third power of the linear dimensions. The variation of current with the square of linear dimensions is evident when it is considered that any increase in dimensions permits a corresponding increase in belt width and also a corresponding increase in the number of belts which can be introduced to operate in parallel.

In adapting the design of Fig. 1 to operate in some other medium, such as in a vacuum or in a liquid, it is only necessary to replace the brush discharge method of introducing charge to and from the belt, by some other charging and discharging device appropriate to the medium. Under such conditions

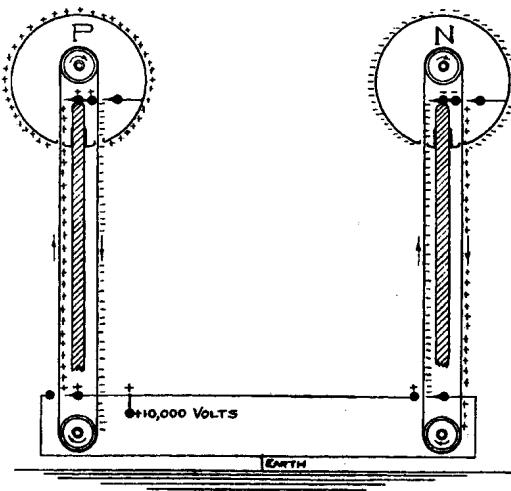


Figure 2:

it may also be advantageous to construct the belt of alternate conducting and insulating segments. It is obvious that the current output of this device can be doubled if the belt on its passage out of the sphere can carry away

a charge equal and opposite to that brought in by the incoming belt. This can be realized very simply through the separation of induced charges by the arrangement which is shown schematically in Fig. 2. This exemplifies the first refinement referred to in the preceding section. The second refinement there mentioned is illustrated in Fig. 3, in which the transformerkenetron set which charges the belts is omitted and the connections are made in

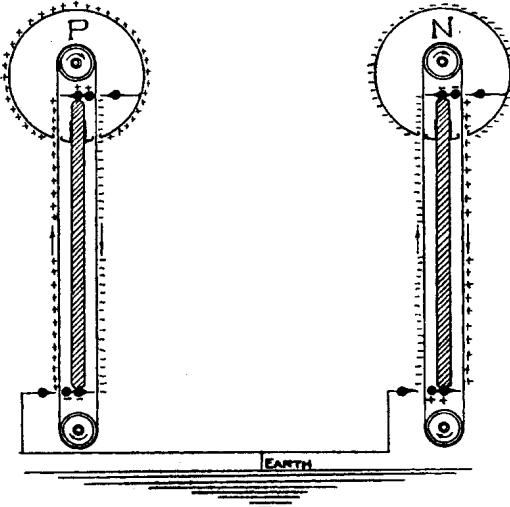


Figure 3:

such a way that a small charge on the moving belt, such as is present due to friction of the belts and pulleys, results in the cumulative separation of positive and negative electricity by induction, thus priming the machine, which immediately begins to operate self-exciting and at full power, in this way dispensing with the necessity of any external electrical connection.

#### 4 Preliminary Models of the Generator

(1) In 1929 a small model was constructed and used to demonstrate the soundness of the principles involved. This model was hastily improvised but performed as expected. The highest voltage obtained was about 80,000 volts, being limited by the electrical breakdown of the surrounding air.

(2) Following this there was constructed a larger generator contained in a dismountable tank which could be highly evacuated by a suitably designed

combination of high-speed mercury condensation pump and liquid air trap. This type of generator is still in the stage of development but has operated in vacuum at 50,000 volts. Progress has been slow due to the necessity of developing certain necessary points of vacuum technique, and also due to the fact that this work had to be temporarily discontinued for the past eighteen months. However, no unexpected difficulties have been encountered and it is hoped that this generator will be in successful operation in a few months. Thus far experience has confirmed our confidence that vacuum insulation is the ultimate insulation for electrostatic devices and will open up tremendous possibilities in this field.

(3) When it became evident that considerable work would still be required to perfect the vacuum generator, construction was begun in June, 1931 of a generator operating in air, powerful enough to be of scientific use and to give a demonstration of the possibilities of generators of this type. This generator was designed for and developed about 1,500,000 volts and delivered a current of about 25 microamperes. It was constructed with 24-inch spherical electrodes mounted on 7-foot upright Pyrex rods and charged by 2.2-inch silk ribbon belts moving with a linear speed of 3500-foot per minute, and it operated either by self-excitation or by the spraying on of charge from a small 10,000-volt transformer kenotron set. It is interesting to note that although it was constructed at a total cost for materials of only about \$100 it developed approximately twice the voltage of any previous direct-current source of which we have knowledge. This generator was described<sup>10</sup> at the Schenectady meeting of the American Physical Society in September, 1931.

### Current output

If a charge is uniformly distributed on the surface of a linear strip in a region otherwise free from electrical forces, there is set up a field perpendicular to the strip and proportional to the surface density of charge as given by Coulomb's law. The maximum charge which can be held by the strip is that which gives rise to a field equal to the breakdown strength of the surrounding insulating medium. In the case of air this limiting field is about 30,000 volts per centimeter, from which we calculate that the maximum charge on a belt in air amounts to  $2.65 \times 10^{-9}$  coulombs per sq. cm on each side of the belt. This figure when multiplied by the number of sq. cm of belt surface which enter the sphere per second, gives the maximum possible

---

<sup>10</sup>Abstract, Phys. Rev. **38**, 1919 (1931).

current output in amperes. Of course, this current output will be doubled if the belt leaves the sphere also fully charged, but with electricity of opposite sign. In the generator described above as preliminary model 3, the actual current output amounted to about one-fourth of this theoretical maximum. The reasons for this discrepancy were known at the time, but were neglected in the construction of this demonstration model. By more careful design, the output may be brought closer to the theoretical limit, as more recent tests have shown.

There are several factors which account for this inability to attain ideally maximum current output. (1) Only one side of the belt is "sprayed" with charge. This may be overcome by using, for example, a double layer belt split by an earthed metal separator between two sets of converging spraying points and traversing the rest of its path simply as a 2-ply belt charged on both outside surfaces. (2) There are inevitable irregularities in the surface, and the breakdown of insulation in those regions which are electrically overstressed, results in a diminution of the charge carried in those localities. (3) There is an additional component to the electric field arising from the difference in potential between the sphere and the earth, so that the electric intensity at any point is the vector sum of this field and that arising from the charge of the belt. A nonlinear potential distribution along the belt may result in overstressing of the insulating medium in the region just outside the sphere, resulting in failure of the belt to retain its maximum charge while traversing that region. This difficulty is obviated by the use of a supplementary device for insuring uniform potential distribution, as described below. (4) Of course the charging device must be adequate to supply the charge. With a wide belt, a multiplicity of spraying points may be necessary to insure complete and adequate coverage. (5) If the belt is charged by a brush discharge, as in this model, the surrounding air is of course partially ionized, and this condition tends also to reduce the charge which can be placed on the belt by reducing the breakdown strength of the air in the charging locality to some value less than 30,000 volts per cm. This limitation could be removed by use of a non-ionizing device for charging the belts. The brush discharge method of charging, however, has the great advantage of simplicity.

It is impossible to increase the charge carrying capacity of the belt by any change in distribution of charge within the belt, such as by the substitution of volume charges in a moving element for surface charges. This is due to the fact that the limit to the net charge is set by the field just outside the surface of the moving element and is thus independent of the distribution of charges within it.

A consideration of the nature of the limiting electric field around the belts shows that the ascending and descending belts may be placed as close together as desired without reducing their current carrying capacity, the limitation being set only by mechanical considerations such as friction. For this reason it is possible to introduce multiple belts, alternately ascending and descending and packed very closely together and so to increase the current output to quite large values.

### **Efficiency**

The work involved in operating this generator is consumed in overcoming the friction of the pulleys and the resistance to the motion of the belt in the surrounding medium, and in the transference of electric charge from earth potential to the potential of the spherical terminals. As is well known, a belt is one of the most efficient means of transferring power, and the two inherent types of electrical losses may be reduced to very small amounts. The first of these arises from electrical leakage, which may be controlled and reduced by methods described in the next section. The second is loss in the process of charging and discharging the belts, which may be reduced to that corresponding to the voltage required to maintain corona discharge from the spraying and discharging points, a voltage which is insignificant in comparison with the voltage generated. There are of course no magnetic losses. Thus this type of generator should be capable of operation with high efficiency.

### **Disturbing factors**

The only disturbing factor which has been found to affect the satisfactory operation of the generator is electrical leakage which is closely identified with two factors, humidity and geometrical design of insulating support. The difficulty is not so much from direct electrical loss by leakage over the surface of the support, as from distortion of the electric field about the spheres by the charges which leak down the supporting insulator and in this way promote insulation breakdown of the air surrounding the insulator near where it enters the sphere. This disturbing factor is completely eliminated by proper design of the insulating support, consisting in the substitution of a hollow insulating cylinder for the insulating rod of model 3. The interior of this cylinder can be maintained at low relative humidity by warming the air within it, so that its interior and exterior surfaces, as well as the belts which run within it, are maintained in the most favorable conditions for

elimination of leakage. In addition to improving insulation, the cylindrical support introduces increased mechanical strength, quietness of operation and safety, and certain other advantages.

A number of important advantages are secured by introducing on the surface of the supporting cylinder an artificial leak from the sphere to the ground, so constructed as to give the most favorable distribution of field between the sphere and the earth. This artificial leak may be constructed, for example, by rotating the cylinder in a lathe and drawing on its surface a continuous, closely spaced, helical, India ink line, extending from one end to the other. By means of this artificial leak the vertical field between the earth and the sphere may be made uniform, thus minimizing leakage to earth through the support or the surrounding air, and permitting the maximum charge to be carried by the belt throughout its entire path.

## 5 The Large Generator Under Construction at Round Hill

The favorable experience with the preliminary models described above appeared to justify the construction of a generator capable of yielding the maximum performance which can reasonably be expected in a generator operating in air, subject to limitations in voltage set by the size of the building in which it is placed. The generator was therefore designed to take full advantage of the largest laboratory space available, which was the airship dock on the estate of Colonel E. H. R. Green, kindly put at our disposal for this purpose. This dock is a building of structural steel covered by corrugated sheet metal and has dimensions approximately  $140 \times 75 \times 75$  ft. In the back end of this building a row of low rooms has been erected to serve as shop and office headquarters, while running lengthwise down the middle of the floor space there has been installed a railroad track of 14-ft. gauge, extending through the huge doors at the front of the building out into the open air for a distance of 160 ft. On this track each of the two generating units is mounted on a truck of structural steel, so that their distance may be varied and they may be run out-of-doors for experiments in the open air. In this connection it is interesting to note that Round Hill is extremely exposed to fogs from the ocean, so that the experience with this generator will afford a severe test of the utility of an electrostatic device under adverse conditions.

A schematic view of the large generator is given by Fig. 4. Its detailed description will be postponed to include its performance tests, but the fol-

lowing information may be of interest.

The spheres are of aluminum alloy 15 ft. in diameter with walls 1/4-inch thick. They were pressed and shipped in "orange peel" sections which were then welded together, and the outer surface ground and polished. Each sphere has four circular holes; a six foot hole on the bottom admits the multiple belts; a trap door on the lower side permits entrance via a ladder from the ground; a trap door on top permits access to the top of the sphere; a trap door on the equator will admit the end of a large discharge tube, spanning the gap between the spheres, to project within the sphere for attachment of subsidiary apparatus and connections and for operation. Preliminary tests on such a discharge tube have been made<sup>11</sup>. The inside of each sphere is itself a laboratory room, provided with a floor and containing accessory apparatus.

The insulating supports are cylinders of Textolite about 24 ft. high, 6 ft. in diameter and of 5/8-inch wall thickness. Each cylinder consists of three 8-foot sections, joined by internal Textolite bands fastened with Textolite dowel pins. Thus the external surface is smooth for easy application of the artificial surface leak and there is an absence of intermediate metal parts, as their presence would distort the field.

The trucks are made of structural steel and are so designed as to permit easy access to or change of the assembly of motors, pulleys and belts.

These three elements, sphere, insulating support and truck, are fundamental elements of any high-voltage assembly, and permit complete freedom for future experimentation and further development of the devices for generation. At the time of writing the spheres are built and polished, the trucks are completed and the main structure is now being assembled. The spheres were supplied by the Chicago Bridge and Iron Works and the Textolite cylinders (Shellac Compound, N. 974) by the General Electric Company, both of which organizations have been very cooperative in their efforts to make the generator successful.

On the basis both of theory and of past experience, we expect this generator to develop about 10,000,000 volts. The power output will depend on the number, size and speed of the belts. Present plans are for an output of about 20 kw. This could be greatly increased by the installation of additional belts, but for the initial adjustments only a portion of this power will be required.

Without the cooperation of many people this project could not have been carried through to its present state. It is impossible to make all the

---

<sup>11</sup>Van Atta, Van de Graaff and Barton, Phys. Rev. **43**, 158 (1933).

acknowledgments which are due, but we cannot let the opportunity pass for expressing the following acknowledgments.

We wish to express our great indebtedness to the Research Corporation

ELECTROSTATIC GENERATORS

157

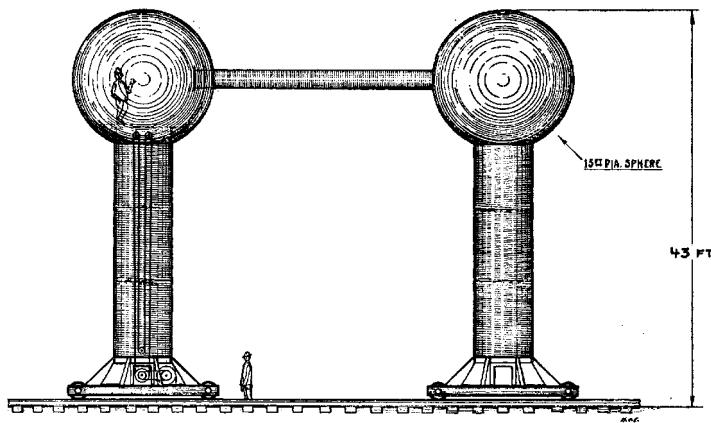


Figure 4:

of New York for a grant defraying a considerable portion of the expenses involved in the construction of this large generator, and for the great interest and help which its officers have given in the development of the engineering plans and the carrying through of the entire project up to the point of assembly.

This new high-voltage installation will be a significant addition to the facilities for scientific and engineering research in the Experiment Station of the Massachusetts Institute of Technology which is operated, with the generous cooperation and support of Colonel E. H. R. Green, on his estate at Round Hill. We wish particularly to thank him for his permission to transform the airship dock, equipped with electric power and other services, into a high-voltage laboratory. We are also greatly indebted to the New York, New Haven and Hartford Railroad for their donation of the railroad track.

Many of our colleagues have been very kind in their assistance. We wish particularly to mention Professor E. L. Bowles who is in supervisory charge of M. I. T. operations at Round Hill.

We wish finally to acknowledge the kind cooperation of Princeton University in whose Palmer Physical Laboratory the preliminary models of this

generator were built, and which sponsored the project up to the time when the new facilities at Round Hill became available.

## The Positive Electron

CARL D. ANDERSON,  
California Institute of Technology, Pasadena, California  
(Received February 28, 1933)

### Abstract

Out of a group of 1300 photographs of cosmic-ray tracks in a vertical Wilson chamber 15 tracks were of positive particles which could not have a mass as great as that of the proton. From an examination of the energy-loss and ionization produced it is concluded that the charge is less than twice, and is probably exactly equal to, that of the proton. If these particles carry unit positive charge the curvatures and ionizations produced require the mass to be less than twenty times the electron mass. These particles will be called positrons. Because they occur in groups associated with other tracks it is concluded that they must be secondary particles ejected from atomic nuclei.

Editor

On August 2, 1932, during the course of photographing cosmic-ray tracks produced in a vertical Wilson chamber (magnetic field of 15,000 gauss) designed in the summer of 1930 by Professor R. A. Millikan and the writer, the tracks shown in Fig. 1 were obtained, which seemed to be interpretable only on the basis of the existence in this case of a particle carrying a positive charge but having a mass of the same order of magnitude as that normally possessed by a free negative electron. Later study of the photograph by a whole group of men of the Norman Bridge Laboratory only tended to strengthen this view. The reason that this interpretation seemed so inevitable is that the track appearing on the upper half of the figure cannot possibly have a mass as large as that of a proton for as soon as the mass is fixed the energy is at once fixed by the curvature. The energy of a proton of that curvature comes out 300,000 volts, but a proton of that energy accord-

ing to well established and universally accepted determinations<sup>1</sup> has a total range of about 5 mm in air while that portion of the range actually visible in this case exceeds 5 cm without a noticeable change in curvature. The only escape from this conclusion would be to assume that at exactly the same instant (and the sharpness of the tracks determines that instant to within about a fiftieth of a second) two independent electrons happened to produce two tracks so placed as to give the impression of a single particle shooting through the lead plate. This assumption was dismissed on a probability basis, since a sharp track of this order of curvature under the experimental conditions prevailing occurred in the chamber only once in some 500 exposures, and since there was practically no chance at all that two such tracks should line up in this way. We also discarded as completely untenable the assumption of an electron of 20 million volts entering the lead on one side and coming out with an energy of 60 million volts on the other side. A fourth possibility is that a photon, entering the lead from above, knocked out of the nucleus of a lead atom two particles, one of which shot upward and the other downward. But in this case the upward moving one would be a positive of small mass so that either of the two possibilities leads to the existence of the positive electron.

In the course of the next few weeks other photographs were obtained which could be interpreted logically only on the positive-electron basis, and a brief report was then published<sup>2</sup> with due reserve in interpretation in view of the importance and striking nature of the announcement.

## 1 MAGNITUDE OF CHARGE AND MASS

It is possible with the present experimental data only to assign rather wide limits to the magnitude of the charge and mass of the particle. The specific ionization was not in these cases measured, but it appears very probable, from a knowledge of the experimental conditions and by comparison with many other photographs of high- and low-speed electrons taken under the same conditions, that the charge cannot differ in magnitude from that of an electron by an amount as great as a factor of two. Furthermore, if the photograph is taken to represent a positive particle penetrating the 6 mm lead plate, then the energy lost, calculated for unit charge, is approximately

---

<sup>1</sup>Rutherford, Chadwick and Ellis, *Radiations from Radioactive Substances*, p. 294.  
Assuming  $R\alpha v^3$  and using data there given the range of a 300,000 volt proton in air S.T.P. is about 5 mm.

<sup>2</sup>C. D. Anderson, Science **76**, 238 (1932).

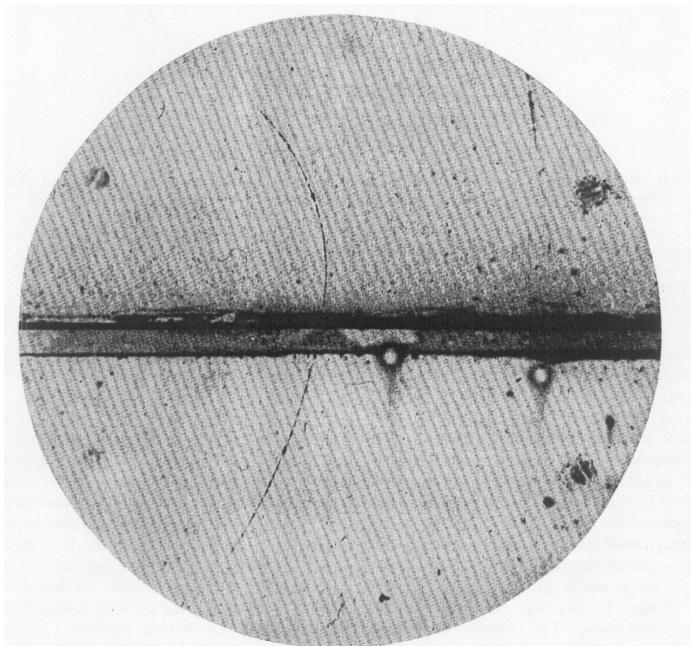


Figure 1: A 63 million volt positron ( $H\rho = 2.1 \times 10^5$  gauss-cm) passing through a 6 mm lead plate and emerging as a 23 million volt positron ( $H\rho = 7.5 \times 10^4$  gauss-cm). The length of this latter path is at least ten times greater than the possible length of a proton path of this curvature.

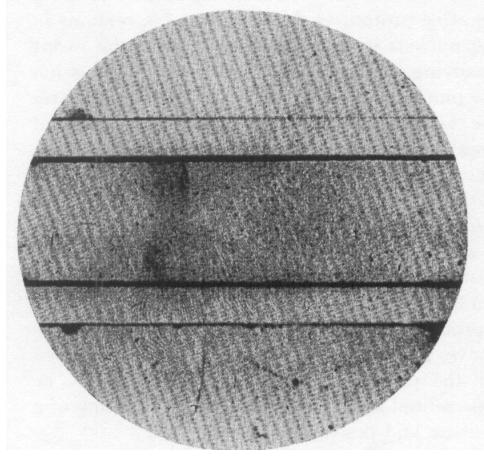


Figure 2: A positron of 20 million volts energy ( $H\rho = 7.1 \times 10^4$  gauss-cm) and a negatron of 30 million volts energy ( $H\rho = 10.2 \times 10^4$  gauss-cm) projected from a plate of lead. The range of the positive particle precludes the possibility of ascribing it to a proton of the observed curvature.

38 million electron-volts, this value being practically independent of the proper mass of the particle as long as it is not too many times larger than that of a free negative electron. This value of 63 million volts per cm energy-loss for the positive particle it was considered legitimate to compare with the measured mean of approximately 35 million volts<sup>3</sup> for negative electrons of 200-300 million volts energy since the rate of energy-loss for particles of small mass is expected to change only very slowly over an energy range extending from several million to several hundred million volts. Allowance being made for experimental uncertainties, an upper limit to the rate of loss of energy for the positive particle can then be set at less than four times that for an electron, thus fixing, by the usual relation between rate of ionization and charge, an upper limit to the charge less than twice that of the negative electron. It is concluded, therefore, that the magnitude of the charge of the positive electron which we shall henceforth contract to positron is very probably equal to that of a free negative electron which from symmetry considerations would naturally then be called a negatron.

It is pointed out that the effective depth of the chamber in the line of sight which is the same as the direction of the magnetic lines of force was 1

---

<sup>3</sup>C. D. Anderson, Phys. Rev. **43**, 381A (1933).

cm and its effective diameter at right angles to that line 14 cm, thus insuring that the particle crossed the chamber practically normal to the lines of force. The change in direction due to scattering in the lead<sup>4</sup>, in this case about 8° measured in the plane of the chamber, is a probable value for a particle of this energy though less than the most probable value.

The magnitude of the proper mass cannot as yet be given further than to fix an upper limit to it about twenty times that of the electron mass. If Fig. 1 represents a particle of unit charge passing through the lead plate then the curvatures, on the basis of the information at hand on ionization, give too low a value for the energy-loss unless the mass is taken less than twenty times that of the negative electron mass. Further determinations of  $H\rho$  for relatively low energy particles before and after they cross a known amount of matter, together with a study of ballistic effects such as close encounters with electrons, involving large energy transfers, will enable closer limits to be assigned to the mass.

To date, out of a group of 1300 photographs of cosmic-ray tracks 15 of these show positive particles penetrating the lead, none of which can be ascribed to particles with a mass as large as that of a proton, thus establishing the existence of positive particles of unit charge and of mass small compared to that of a proton. In many other cases due either to the short section of track available for measurement or to the high energy of the particle it is not possible to differentiate with certainty between protons and positrons. A comparison of the six or seven hundred positive-ray tracks which we have taken is, however, still consistent with the view that the positive particle which is knocked out of the nucleus by the incoming primary cosmic ray is in many cases a proton.

From the fact that positrons occur in groups associated with other tracks it is concluded that they must be secondary particles ejected from an atomic nucleus. If we retain the view that a nucleus consists of protons and neutrons (and  $\alpha$ - particles) and that a neutron represents a close combination of a proton and electron, then from the electromagnetic theory as to the origin of mass the simplest assumption would seem to be that an encounter between the incoming primary ray and a proton may take place in such a way as to expand the diameter of the proton to the same value as that possessed by the negatron. This process would release an energy of a billion electron-volts appearing as a secondary photon. As a second possibility the primary ray may disintegrate a neutron (or more than one) in the nucleus by the ejection either of a negatron or a positron with the result that a positive or

---

<sup>4</sup>C. D. Anderson, Phys. Rev. **43**, 381A (1933).

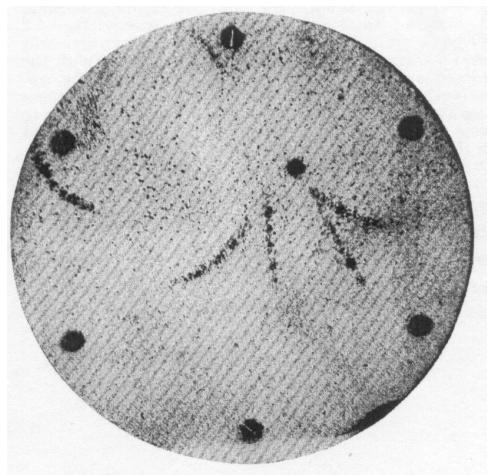


Figure 3: A group of six particles projected from a region in the wall of the chamber. The track at the left of the central group of four tracks is a negatron of about 18 million volts energy ( $H\rho = 6.2 \times 10^4$  gauss-cm) and that at the right a positron of about 20 million volts energy ( $H\rho = 7.0 \times 10^4$  gauss-cm). Identification of the two tracks in the center is not possible. A negatron of about 15 million volts is shown at the left. This group represents early tracks which were broadened by the diffusion of the ions. The uniformity of this broadening for all the tracks shows that the particles entered the chamber at the same time.

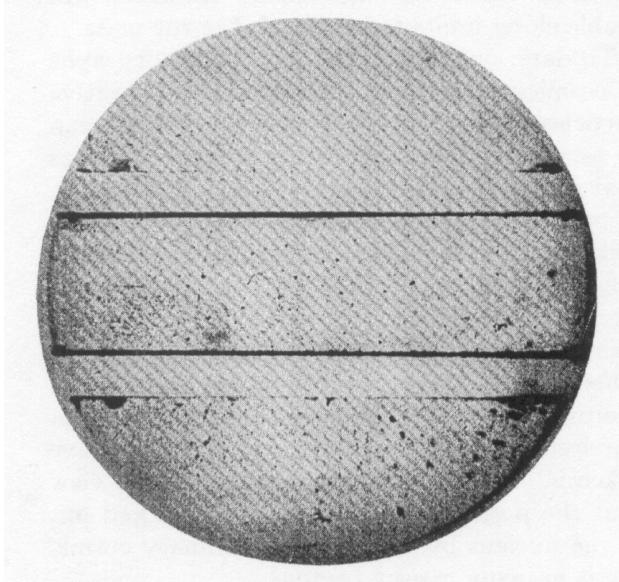


Figure 4: A positron of about 200 million volts energy ( $H\rho = 6.6 \times 10^5$  gauss-cm) penetrates the 11 mm lead plate and emerges with about 125 million volts energy ( $H\rho = 4.2 \times 10^5$  gauss-cm). The assumption that the tracks represent a proton traversing the lead plate is inconsistent with the observed curvatures. The energies would then be, respectively, about 20 million and 8 million volts above and below the lead, energies too low to permit the proton to have a range sufficient to penetrate a plate of lead of 11 mm thickness.

a negative proton, as the case may be, remains in the nucleus in place of the neutron, the event occurring in this instance without the emission of a photon. This alternative, however, postulates the existence in the nucleus of a proton of negative charge, no evidence for which exists. The greater symmetry, however, between the positive and negative charges revealed by the discovery of the positron should prove a stimulus to search for evidence of the existence of negative protons. If the neutron should prove to be a fundamental particle of a new kind rather than a proton and negatron in close combination, the above hypotheses will have to be abandoned for the proton will then in all probability be represented as a complex particle consisting of a neutron and positron.

While this paper was in preparation press reports have announced that P. M. S. Blackett and G. Occhialini in an extensive study of cosmic-ray tracks have also obtained evidence for the existence of light positive particles confirming our earlier report.

I wish to express my great indebtedness to Professor R. A. Millikan for suggesting this research and for many helpful discussions during its progress. The able assistance of Mr. Seth H. Neddermeyer is also appreciated.

THE HIGHLY COLLAPSED CONFIGURATIONS OF A  
STELLAR MASS. (SECOND PAPER.)

*S. Chandrasekhar, Ph.D.*

1. A study of the equilibrium of degenerate gas spheres has a twofold significance in the analysis of stellar structure, namely, in providing an approach to a proper theory of white dwarfs, and also, we shall see, in providing a certain limiting sequence of configurations to which all stars must tend eventually. A beginning in the study of these configurations was made by the author in a previous communication,\* where for convenience the equation of state of degenerate matter was taken to correspond to one or other of the two limiting forms  $p = K_1 \rho^{5/3}$  or  $p = K_2 \rho^{4/3}$  according as the density was less than or greater than a certain density  $\rho'$  where

$$\rho' = (K_2/K_1)^{3/2},$$

$\rho'$  itself being such that both the equations of state yield the same calculated value for the pressure. Actually in the analysis a certain small temperature gradient was allowed for. Working on the standard model it was assumed that the ratio  $\beta$  of the gas pressure to the total pressure was a constant, but by hypothesis ("highly collapsed")  $\beta$  was taken to be very nearly unity. On these assumptions it followed that stars of mass less than a certain specified  $M_{3/2}$  (see I, § 6, page 462) were complete Emden polytropes with index  $n = 3/2$ , and further that configurations of greater mass must be *composite*, *i.e.* must have inner regions where degeneracy is predominantly relativistic. Lastly, and this was the most important conclusion reached, these composite configurations have a *natural limit*: On the standard model a completely relativistically degenerate configuration has a mass given by (*cf.* I, equation (36))

$$M = -\frac{4}{\pi^{1/2}} \left( \frac{K_2}{G} \right)^{3/2} \left( \xi^2 \frac{d\theta_3}{d\xi} \right) \cdot \beta^{-3/2} = M_3 \beta^{-3/2} \text{ (say),} \quad (1)\dagger$$

where  $\theta_3$  is the Emden function with index  $n = 3$ . These configurations have zero radius (*cf.* the remarks in I following the equations (45), (46), page 463).‡

\* *M.N.*, 91, 456, 1931 (referred to as I). See also the earlier papers of the author in *Phil. Mag.*, 11, 592, 1931, and *Astrophysical Journal*, 64, 92.

† In I we denoted by  $M_3$  what we have now defined as  $M_3 \beta^{-3/2}$ . It is convenient to separate out the term involving  $\beta$  from the purely "mass factor."

‡ In I this "singularity" was formally avoided by introducing a state of "maximum density" for matter, but now we shall not introduce any such hypothetical states, mainly for the reason that it appears from general considerations that when the central density is high enough for marked deviations from the known gas laws (degenerate or otherwise) to occur the configurations then would have such small radii that they would cease to have any practical importance in astrophysics.

Apart from the above results of a general character, the analysis in I did not lead to any further quantitative results. To obtain by the methods of I anything more exact would have meant very considerable numerical work to "fit" an appropriate solution of Emden's equation with index  $n=3/2$  (to describe the outer ordinarily degenerate envelope) with an *Emden function* of index 3 (to describe the inner relativistically degenerate core). It would be very much more satisfactory to take the exact equation describing the degenerate state and treat the whole degenerate parts of a star on the same footing instead of as in I, further subdividing it to correspond to one or other of the two limiting forms of the equation describing the degenerate state. By a very remarkable coincidence the differential equation (governing the structure of a degenerate gas sphere in hydrostatic equilibrium) based on the exact equation of state takes an extremely simple form. We show, in fact, that the structure of the configuration is governed by a solution of the differential equation,

$$\frac{1}{\eta^2} \frac{d}{d\eta} \left( \eta^2 \frac{d\phi}{d\eta} \right) = - \left( \phi^2 - \frac{1}{y_0^2} \right)^{3/2}. \quad (2)*$$

It is to be noticed that there is only one parameter occurring in the equation, and a single system of integrations should suffice to obtain a clear insight into these configurations. Equation (2) has a formal similarity with Emden's equation. Indeed, we shall show that under certain circumstances  $\phi$  can be expressed in terms of the Emden functions with appropriate indices. It is the derivation of the above equation that has led to the developments summarised in this and the following paper. In this paper we shall establish this equation and provide tables of solutions. In the analysis we shall omit all references to radiation pressure, *i.e.* this paper strictly deals with configurations having  $\beta=1$ . The introduction of radiation in these configurations involves quite delicate considerations, and all these find a proper treatment in the paper following this one.

*2. The Differential Equation governing the Structure of Degenerate Matter in Gravitational Equilibrium.*—The pressure-density relation for a degenerate gas can be written parametrically as follows:—

$$\left. \begin{aligned} p &= \frac{\pi m^4 c^5}{3h^3} [x(2x^2 - 3)(x^2 + 1)^{1/2} + 3 \sinh^{-1} x], \\ \rho &= \frac{8\pi m^3 c^3 \mu H}{3h^3} x^3, \end{aligned} \right\} \quad (3)$$

where  $m$  = mass of the electron,  $c$  = velocity of light,  $h$  = Planck's constant,  $H$  = mass of the proton,  $\mu$  = molecular weight. Equation (3) is established in Appendix I to this paper, where also  $f(x)$  is tabulated. We rewrite (3) as

$$p = A_2 f(x); \quad \rho = B x^3, \quad (4)$$

---

\* This equation was given without proof in the author's preliminary note in the *Observatory*, 57, 373, 1934.

where

$$\left. \begin{aligned} A_2 &= \frac{\pi m^4 c^5}{3 h^3}; & B &= \frac{8\pi m^3 c^3 \mu H}{3 h^3}, \\ f(x) &= x(2x^2 - 3)(x^2 + 1)^{1/2} + 3 \sinh^{-1} x. \end{aligned} \right\} \quad (5)$$

The equations of equilibrium are, as usual,

$$\left. \begin{aligned} \frac{dp}{dr} &= -\frac{GM(r)}{r^2}\rho, \\ \frac{dM(r)}{dr} &= 4\pi\rho r^2. \end{aligned} \right\} \quad (6)$$

From (6) we have

$$\frac{1}{r^2} \frac{d}{dr} \left( \frac{r^2}{\rho} \frac{dp}{dr} \right) = -4\pi G\rho. \quad (7)$$

Substitute for  $p$  and  $\rho$  from (4). We have

$$\frac{A_2}{B} \frac{1}{r^2} \frac{d}{dr} \left( \frac{r^2}{x^3} \frac{df(x)}{dr} \right) = -4\pi GBx^3. \quad (8)$$

From the definition of  $f(x)$  in (5) we easily verify that

$$\frac{df(x)}{dx} = \frac{8x^4}{(x^2 + 1)^{1/2}} \frac{dx}{dr}, \quad (9)$$

or

$$\frac{1}{x^3} \frac{df(x)}{dr} = \frac{8x}{(x^2 + 1)^{1/2}} \frac{dx}{dr} = 8 \frac{d\sqrt{x^2 + 1}}{dr}. \quad (10)$$

Hence (8) can be rewritten as

$$\frac{1}{r^2} \frac{d}{dr} \left( r^2 \frac{d\sqrt{x^2 + 1}}{dr} \right) = -\frac{\pi GB^2}{2A_2} x^3. \quad (11)$$

Put

$$y^2 = x^2 + 1. \quad (12)$$

Then

$$\frac{1}{r^2} \frac{d}{dr} \left( r^2 \frac{dy}{dr} \right) = -\frac{\pi GB^2}{2A_2} (y^2 - 1)^{3/2}. \quad (13)$$

Let  $x$  take the value  $x_0$  at the centre.

Further, let  $y_0$  be the corresponding value of  $y$  at the centre. Introduce the new variables  $\eta$  and  $\phi$  defined as follows :—

$$r = a\eta; \quad y = y_0\phi, \quad (14)$$

where

$$\left. \begin{aligned} a &= \left( \frac{2A_2}{\pi G} \right)^{1/2} \frac{1}{By_0}, \\ y_0^2 &= x_0^2 + 1. \end{aligned} \right\} \quad (15)$$

Our differential equation finally takes the form

$$\frac{1}{\eta^2} \frac{d}{d\eta} \left( \eta^2 \frac{d\phi}{d\eta} \right) = -\left( \phi^2 - \frac{1}{y_0^2} \right)^{3/2}. \quad (16)$$

By (14) we have to seek a solution of (16) such that  $\phi$  takes the value unity at the origin. Further, from symmetry the derivative of  $\phi$  must

vanish at the origin. The *boundary* is defined at the point where the density vanishes, and this by (12) means that if  $\eta_1$  specifies the boundary

$$\phi(\eta_1) = \frac{1}{y_0}. \quad (17)$$

3. From our definitions of the various quantities we find that

$$\rho = \rho_0 \frac{y_0^3}{(y_0^2 - 1)^{3/2}} \left( \phi^2 - \frac{1}{y_0^2} \right)^{3/2}, \quad (18)$$

where

$$\rho_0 = Bx_0^3 = B(y_0^2 - 1)^{3/2} \quad (18')$$

specifies the central density. Also we may notice that the scale of length  $a$  introduced in (15) has in terms of the physical quantities the form

$$a = \frac{1}{4\pi m \mu H y_0} \left( \frac{3h^3}{2cG} \right)^{1/2}, \quad (19)$$

or putting in numerical values

$$a = \frac{7.720 \times 10^8}{\mu y_0} = l_1 y_0^{-1} \text{ cm. (say)}. \quad (20)$$

4. *The Potential.*—The function  $\phi$  itself has a physical meaning. If  $V$  is the inner gravitational potential, then from general theory we have

$$\frac{dV}{dr} = \frac{1}{\rho} \frac{dP}{dr}. \quad (21)$$

From (5) and (10) we see that

$$\frac{dV}{dr} = \frac{8A_2}{B} y_0 \frac{d\phi}{dr}, \quad (22)$$

or integrating

$$V = \frac{8A_2}{B} y_0 \phi + \text{constant}. \quad (23)$$

If we choose the arbitrary zero of the potential on the boundary of the configuration we have by (17) that the “constant” in (23) is  $(8A_2/B)$ . Hence finally

$$V = \frac{8A_2}{B} y_0 \left( \phi - \frac{1}{y_0} \right). \quad (24)$$

5. *The Mass Relation.*—The mass of the material enclosed up to a point  $\eta$  is clearly

$$M(\eta) = 4\pi \int_0^\eta \rho r^2 dr = 4\pi a^3 \int_0^\eta \rho \eta^2 d\eta. \quad (25)$$

By (18),

$$M(\eta) = 4\pi \rho_0 \frac{a^3 y_0^3}{(y_0^2 - 1)^{3/2}} \int_0^\eta \left( \phi^2 - \frac{1}{y_0^2} \right)^{3/2} \eta^2 d\eta, \quad (26)$$

or using our differential equation (16)

$$M(\eta) = -4\pi \rho_0 \frac{a^3 y_0^3}{(y_0^2 - 1)^{3/2}} \int_0^\eta \frac{d}{d\eta} \left( \eta^2 \frac{d\phi}{d\eta} \right) d\eta. \quad (27)$$

Remembering that  $\rho_0$  is given by (18) we have explicitly

$$M(\eta) = -4\pi \left( \frac{2A_2}{\pi G} \right)^{3/2} \frac{1}{B^2} \eta^2 \frac{d\phi}{d\eta}. \quad (28)$$

The mass of the whole configuration is therefore

$$M = -4\pi \left( \frac{2A_2}{\pi G} \right)^{3/2} \frac{1}{B^2} \left( \eta^2 \frac{d\phi}{d\eta} \right)_{\eta=\eta_1}. \quad (29)$$

We notice that in (28) and (29)  $y_0$  does not *explicitly* occur. It is of course implicitly present inasmuch as in the differential equation defining  $\phi$ ,  $y_0$  occurs.

6. *The Relation between the Mean and the Central Density.*—Let  $\bar{\rho}(\eta)$  be the mean density of the material inside  $\eta$ . Then

$$M(\eta) = \frac{4}{3}\pi\alpha^3\eta^3\bar{\rho}(\eta). \quad (30)$$

Comparing (28) and (30), we have

$$\frac{1}{3}\eta^3\bar{\rho}(\eta) = -\rho_0 \frac{y_0^3}{(y_0^2 - 1)^{3/2}} \eta^2 \frac{d\phi}{d\eta}, \quad (31)$$

or

$$\frac{\bar{\rho}(\eta)}{\rho_0} = -3 \frac{y_0^3}{(y_0^2 - 1)^{3/2}} \frac{1}{\eta} \frac{d\phi}{d\eta}. \quad (32)$$

From (32) we deduce that *the relation between the mean and the central density of the whole configuration is*

$$\rho_0 = -\bar{\rho} \left( 1 - \frac{1}{y_0^2} \right)^{3/2} \frac{\eta_1}{3\phi'(\eta_1)} \quad (33)$$

( $\phi'$  denoting the derivative)—a relation analogous to the corresponding relation in the theory of polytropes.

7. *An Approximation for Configurations with Small Central Densities.*—When the central density is small we should have the law  $p = K_1^{5/3}$  holding approximately, and the corresponding configurations must have structures which can approximately be represented by an Emden polytrope with index  $n = 3/2$ . We establish this on our differential equation in the following way:—

Now by definition  $y_0^2 = x_0^2 + 1$ , and we need a first-order approximation when  $x_0^2$  is small. *We shall neglect all quantities of order  $x_0^4$  or higher.* Then

$$y_0 = 1 + \frac{1}{2}x_0^2. \quad (34)$$

Put

$$\phi^2 - \frac{1}{y_0^2} = \theta. \quad (35)$$

In our approximation we have

$$\phi = 1 - \frac{1}{2}(x_0^2 - \theta). \quad (36)$$

At the origin  $\phi$  takes the value unity. Hence

$$\theta(0) = x_0^2. \quad (37)$$

From (16) we derive the following differential equation for  $\theta$  :—

$$\frac{1}{2} \frac{d^2\theta}{d\eta^2} + \frac{1}{\eta} \frac{d\theta}{d\eta} = -\theta^{3/2}. \quad (38)$$

Put

$$\xi = 2^{1/2}\eta. \quad (39)$$

Then

$$\frac{1}{\xi^2} \frac{d}{d\xi} \left( \xi^2 \frac{d\theta}{d\xi} \right) = -\theta^{3/2}, \quad (40)$$

which is Emden's equation with index  $n = 3/2$ , but *the solution we need is not the Emden function in the usual normalisation* \* with  $\theta = 1$  at  $\xi = 0$ . By (37) our  $\theta$  takes the value  $x_0^{-2}$  at the origin. Denote by  $\theta_{3/2}$  the Emden function. Now it is a property of the differential equation (40) that if  $\theta$  is any solution then  $C^4\theta(C\xi)$  is also a solution where  $C$  is any arbitrary constant. Hence if we put

$$C = x_0^{1/2}, \quad (41)$$

and take for  $\theta$ ,  $\theta_{3/2}$ , we would obtain the solution we need. Hence

$$\theta = x_0^{-2}\theta_{3/2}(x_0^{1/2}\xi) = x_0^{-2}\theta_{3/2}(\sqrt{2x_0}\eta). \quad (42)$$

By (37) then

$$\phi = 1 - \frac{1}{2}x_0^{-2}\{1 - \theta_{3/2}(\sqrt{2x_0}\eta)\} + O(x_0^{-4}), \quad (43)$$

which relates  $\phi$  with  $\theta_{3/2}$ . From (43) we see that for these configurations the boundary  $\eta_1$  must be such that

$$(\theta_{3/2}\sqrt{2x_0}\eta_1) = 0. \quad (44)$$

Let  $\xi_1(\theta_{3/2})$  be the boundary of the *Emden function*. Then from (44) we deduce that

$$\eta_1 = \frac{\xi_1(\theta_{3/2})}{\sqrt{2x_0}}. \quad (45)$$

From (45) we see that as  $y_0 \rightarrow 1$ ,  $x_0 \rightarrow 0$ ,  $\eta_1 \rightarrow \infty$ . The radius tends to infinity with the same singularity.

Again from (43) we have

$$\frac{d\phi}{d\eta} = \frac{1}{2}x_0^{-2}\sqrt{2x_0} \frac{d\theta_{3/2}(\xi)}{d\xi}. \quad (46)$$

Combining (45) and (46) we have a relation we shall need later :

$$\left( \eta^2 \frac{d\phi}{d\eta} \right)_1 = \left( \frac{x_0}{2} \right)^{3/2} \left( \xi^2 \frac{d\theta_{3/2}}{d\xi} \right)_1. \quad (47)$$

Further,

$$\left( \frac{1}{\eta} \frac{d\phi}{d\eta} \right)_1 = x_0^3 \left( \frac{1}{\xi} \frac{d\theta_{3/2}}{d\xi} \right)_1. \quad (48)$$

We shall find the above expressions useful when we come to discuss "highly"

\* In the sequel by "Emden function" we shall always mean the one which takes the value unity at the origin. We shall denote the Emden function with index  $n$  by  $\theta_n$ .

collapsed configurations ( $(1 - \beta)$  finite but small), but now we verify that the scheme is consistent. From (48) and (33) we have

$$\rho_0 = -\bar{\rho} \left( \frac{\xi}{3\theta'_{3/2}} \right)_1, \quad (49)$$

which is precisely the formula for an Emden polytrope with index  $n = 3/2$ . Again from (29) and (47)

$$M = -4\pi \left( \frac{2A_2}{\pi G} \right)^{3/2} \frac{1}{B^2} \left( \frac{x_0}{2} \right)^{3/2} \left( \xi^2 \frac{d\theta_{3/2}}{d\xi} \right)_1. \quad (50)$$

To compare the above with the formula derived on the law  $p = K_1 \rho^{5/3}$  we note that the degenerate constant  $K_1$ , given by

$$K_1 = \frac{1}{20} \left( \frac{3}{\pi} \right)^{2/3} \frac{h^2}{m(\mu H)^{5/3}}, \quad (51)$$

is related to our  $A_2$  and  $B$  by the relation

$$K_1 = \frac{8}{5} \frac{A_2}{B^{5/3}}. \quad (52)$$

Combining (50) and (52) and setting  $\lambda_2$  to denote the central density ( $= Bx_0^{-3}$ ) we find that

$$M = -4\pi \left( \frac{5K_1}{8\pi G} \right)^{3/2} \lambda_2^{1/2} \left( \xi^2 \frac{d\theta_{3/2}}{d\xi} \right)_1, \quad (53)$$

which is the usual formula since on the law  $p = K\rho^{1+\frac{1}{n}}$  the polytropic relation is

$$M = -4\pi \left( \frac{(n+1)K}{4\pi G} \right)^{3/2} \lambda_2^{\frac{3-n}{2n}} \left( \xi^2 \frac{d\theta_n}{d\xi} \right)_1. \quad (53')$$

8. *The Limiting Mass.*—From our differential equation (16) we see that

$$\phi \rightarrow \theta_3 \quad \text{as} \quad y_0 \rightarrow \infty. \quad (54)$$

But from (20) we see that at the same time the radius tends to zero. From (28) then

$$M \rightarrow -4\pi \left( \frac{2A_2}{\pi G} \right)^{3/2} \frac{1}{B^2} \left( \xi^2 \frac{d\theta_3}{d\xi} \right)_1. \quad (55)$$

To see that we have now simply recovered our earlier result in I (equation (36)) we have only to notice that the relativistic degenerate constant  $K_2$ , defined by

$$K_2 = \left( \frac{3}{\pi} \right)^{1/3} \frac{hc}{8(\mu H)^{4/3}}, \quad (56)$$

is related to our  $A_2$  and  $B$  by the relation

$$K_2 = \frac{2A_2}{B^{4/3}}. \quad (57)$$

9. As mentioned in § 1, we shall denote by  $M_3$  the mass

$$M_3 = 4\pi \left( \frac{2A_2}{\pi G} \right)^{3/2} \frac{1}{B^2} \omega_3^0, \quad (58)$$

where following Milne we have introduced the quantity  $\omega_3^0$  defined by

$$\omega_3^0 = - \left( \xi^2 \frac{d\theta_3}{d\xi} \right)_1. \quad (59)$$

If we define correspondingly that

$$\Omega(y_0) = - \left( \eta^2 \frac{d\phi}{d\eta} \right)_{\eta=\eta_1} \quad (60)$$

for our function  $\phi$ , then the mass relation can be written as

$$M(y_0)\omega_3^0 = M_3\Omega(y_0). \quad (61)$$

As the mass of the configuration increases monotonically with increasing  $y_0$ , we have the useful inequality

$$\Omega(y_0) > \omega_3^0 \quad (y_0 \text{ finite}). \quad (62)$$

Finally we may note that the insertion of numerical values in our formula for  $M_3$  yields

$$M_3 = 5.728\mu^{-2} \times \odot, \quad (63)$$

where  $\odot$  represents the mass of the Sun.

10. *The General Results.*—In the previous sections, §§ 7, 8, 9, we have merely related our present treatment with the results obtained in I on the basis of the polytropic theory. Those results appear as simple limiting cases. However, the exact treatment on the basis of our differential equation

$$\frac{1}{\eta^2} \frac{d}{d\eta} \left( \eta^2 \frac{d\phi}{d\eta} \right) = - \left( \phi^2 - \frac{1}{y_0^2} \right)^{3/2} \quad (64)$$

at the same time provides much more quantitative information. The boundary conditions

$$\phi = 1, \quad \frac{d\phi}{d\eta} = 0 \quad \text{at} \quad \eta = 0, \quad (65)$$

combined with a particular value for  $y_0$ , would determine  $\phi$  completely, and therefore the mass of the configuration as well. The equation (64) does not admit of a “homology constant,” and hence *each mass has a density distribution characteristic of itself which cannot be inferred from the density distribution in a configuration of a different mass*. This difference between our configurations governed by (64) and polytropes has, as we shall see, an important bearing in the theory of general stellar models considered in the following paper.

Each specified value for  $y_0$  determines uniquely the mass  $M$ , the radius  $R_1$  and the ratio of the mean to the central density. We have (collecting together our earlier results) :

$$M/M_3 = \Omega(y_0)/\omega_3^0, \quad (66)$$

$$R_1/l_1 = \eta_1/y_0, \quad (67)$$

$$\rho_0/B = (y_0^2 - 1)^{3/2}, \quad (68)$$

$$\bar{\rho}/\rho_0 = - \frac{I}{\left(I - \frac{I}{y_0^2}\right)^{3/2}} \frac{3}{\eta_1} \left(\frac{d\phi}{d\eta}\right)_1. \quad (69)$$

In (67) we have introduced a new unit of length ( $l_1 = \alpha y_0$ ),

$$l_1 = \frac{I}{4\pi m \mu H} \left( \frac{3h^3}{2cG} \right) = 7.720 \mu^{-1} \times 10^8 \text{ cm.}, \quad (67')$$

and which therefore does not involve factors in  $y_0$ . Further, the physical variables determining the structure of the configuration are:

$$\rho = \rho_0 \frac{I}{\left(I - \frac{I}{y_0^2}\right)^{3/2}} \left( \phi^2 - \frac{I}{y_0^2} \right)^{3/2}, \quad (70)$$

$$\bar{\rho} = -\rho_0 \frac{I}{\left(I - \frac{I}{y_0^2}\right)^{3/2}} \frac{3}{\eta} \frac{d\phi}{d\eta}, \quad (71)$$

$$M(\eta) \propto -\eta^2 \frac{d\phi}{d\eta}. \quad (72)$$

11. In § 10 we have reduced the problem of the structure of degenerate gas spheres to a study of our functions  $\phi$  for different initially prescribed values for the parameter  $y_0$ . The integration has been numerically effected for the following ten different values of the parameter:—

$$I/y_0^2 = 0.8, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0.05, 0.02, 0.01. \quad (73)$$

The following expansion for  $\phi$  near the origin may be noted here for further reference:—

$$\begin{aligned} \phi = I - \frac{q^3}{6} \xi^2 + \frac{q^4}{40} \xi^4 - \frac{q^5(5q^2 + 14)}{7!} \xi^6 + \frac{q^6(339q^2 + 280)}{3 \times 9!} \xi^8 \\ - \frac{q^7(1425q^4 + 11436q^2 + 4256)}{5 \times 11!} \xi^{10} + \dots \end{aligned} \quad (74)$$

where

$$q^2 = I - \frac{I}{y_0^2}. \quad (75)^*$$

The important quantities of interest are the boundary quantities occurring in equations (66), (67), (69). These are tabulated in Table I for the different values of  $y_0$ .

12. From the figures of Table I it is easy to calculate the mass in units of  $M_3$ , the radius in units of  $l_1$  and the central density ( $= x_0^3$ ) in units of  $B$

\* When  $y_0 \rightarrow \infty$ ,  $q \rightarrow 1$  and the series (74) goes over into the expansion for Emden  $\theta_3$  near the origin (cf. *British Association Tables*, 2, Introduction, equation on top of page v).

( $=9.8848 \times 10^5 \mu$  grams cm. $^{-3}$ ). These express the chief physical characteristics of these configurations in the "natural system" of units occurring in the theory of these configurations. In Table III they are converted into the more conventional system of units expressing the radius and the density in C.G.S. units and the mass in units of the Sun. The actual figures tabulated are for  $\mu = 1$ . The figures for other values of  $\mu$  can be obtained by multiplying  $M$  by  $\mu^{-2}$ ,  $R_1$  by  $\mu^{-1}$  and  $\rho$  by  $\mu$ . To see the order of magnitudes involved here it is of interest to point out that the mass  $4.852\odot\mu^{-2}$  has a radius only slightly over the radius of the Earth (radius of the Earth  $6 \times 10^8$  cm. compared to  $7.7 \times 10^8$  cm. for the radius of  $4.852\odot$ ). The mass  $0.957M_3$  has a radius considerably less than the radius of the Earth.

TABLE I

$\frac{I}{y_0^2}$	$\eta_1$	$-\eta_1^2\phi'(\eta_1)$	$\rho_0/\bar{\rho}$
0	6.8968	2.0182	54.182
.01	5.3571	1.9321	26.203
.02	4.9857	1.8652	21.486
.05	4.4601	1.7096	16.018
.1	4.0690	1.5186	12.626
.2	3.7271	1.2430	9.9348
.3	3.5803	1.0337	8.6673
.4	3.5245	0.8598	7.8886
.5	3.5330	0.7070	7.3505
.6	3.6038	0.5679	6.9504
.8	4.0446	0.3091	6.3814
1	$\infty$	0	5.9907

TABLE II

*The Physical Characteristics of Degenerate Spheres in the "Natural" Units*

$\frac{I}{y_0^2}$	$M/M_3$	$R_1/l_1$	$\rho_0/B$
0	1	0	$\infty$
.01	0.95733	0.53571	985.038
.02	0.92419	0.70508	343
.05	0.84709	0.99732	82.8191
.1	0.75243	1.28674	27
.2	0.61589	1.66682	8
.3	0.51218	1.96102	3.56423
.4	0.42600	2.22908	1.83711
.5	0.35033	2.49818	1
.6	0.28137	2.79148	0.54433
.8	0.15316	3.61760	0.125
1.0	0	$\infty$	0

TABLE III

*The Physical Characteristics of Degenerate Spheres in the Usual Units*(Calculations are for  $\mu = 1$ . For other values  $\mu$ ,  $M$  should be multiplied by  $\mu^{-2}$ ,  $R_1$  by  $\mu^{-1}$ ,  $\rho_c$  by  $\mu$ )

$\frac{I}{y_0'^2}$	$M/\odot$	$\rho_0$ in grm./cm. $^{-3}$	$\rho_{\text{mean}}$ in grm./cm. $^{-3}$	Radius in cm.
0	5.728	$\infty$	$\infty$	0
·01	5.484	$9.737 \times 10^8$	$4.716 \times 10^7$	$4.136 \times 10^8$
·02	5.294	$3.391 \times 10^8$	$1.578 \times 10^7$	$5.443 \times 10^8$
·05	4.852	$8.187 \times 10^7$	$5.111 \times 10^6$	$7.699 \times 10^8$
·1	4.310	$2.669 \times 10^7$	$2.114 \times 10^6$	$9.936 \times 10^8$
·2	3.528	$7.908 \times 10^6$	$7.960 \times 10^5$	$1.287 \times 10^9$
·3	2.934	$3.523 \times 10^6$	$4.065 \times 10^5$	$1.514 \times 10^9$
·4	2.440	$1.816 \times 10^6$	$2.302 \times 10^5$	$1.721 \times 10^9$
·5	2.007	$9.885 \times 10^5$	$1.345 \times 10^5$	$1.929 \times 10^9$
·6	1.612	$5.381 \times 10^5$	$7.741 \times 10^4$	$2.155 \times 10^9$
·8	0.877	$1.236 \times 10^5$	$1.936 \times 10^4$	$2.793 \times 10^9$
1.0	0	0	0	$\infty$

Now if we define that matter is "relativistically degenerate" for densities greater than  $\rho' (= (K_2/K_1)^3)$ , then we can from our results easily find the masses which are characterised by central regions of "relativistic degeneracy." The value of  $x$  corresponding to  $\rho'$  is readily seen to be 1.25. Hence

$$\frac{I}{y_0'^2} = \frac{I}{x'^2 + 1} = 0.39024. \quad (76)$$

From fig. 1 we now see that for  $M \leq 0.43M_3$  there are no regions which are "relativistically degenerate" on this convention. For  $M > 0.43M_3$  there are regions in which  $x > x' (= 1.25)$ , and the fraction of the whole radius inside which  $x > x'$  rapidly increases to unity. In the mass-radius curve we can therefore draw circles about each point with radii proportional to the actual radii of the corresponding configurations, and draw inside each a concentric circle to represent the "relativistic" region. This has been done in fig. 2 at a few points. We see that even for  $M = 0.75M_3$  there is barely a "fringe" of ordinarily degenerate regions. This diagram clearly illustrates a general principle that degeneracy never usually sets in without being relativistic.

13. *Comparison with the Results on Emden Polytrope  $n=3/2$ .*—It is of interest to see in how far the results of the above exact treatment differ from what one would obtain on the law  $p = K_1 \rho^{5/3}$ . We have already shown in § 7 that one gets these Emden configurations as limiting cases for zero density and therefore for small masses (expressed in units of  $M_3$ ). Our comparison here therefore amounts to a comparison of the results based on an exact treatment of the equation (64) with the limiting form for  $y_0 \rightarrow 1$  extrapolated for all masses. For this purpose it is convenient to rewrite the formulæ for the case of the polytrope  $n=3/2$  in the following way.

From (45) and (50) we have now

$$R_1 = \frac{l_1 \xi_1(\theta_{3/2})}{\sqrt{2x_0}}, \quad (77)$$

$$M/M_3 = \left( \frac{x_0}{2} \right)^{3/2} \frac{I}{\omega_3^0} \left( \xi^2 \frac{d\theta_{3/2}}{d\xi} \right)_1. \quad (77')$$

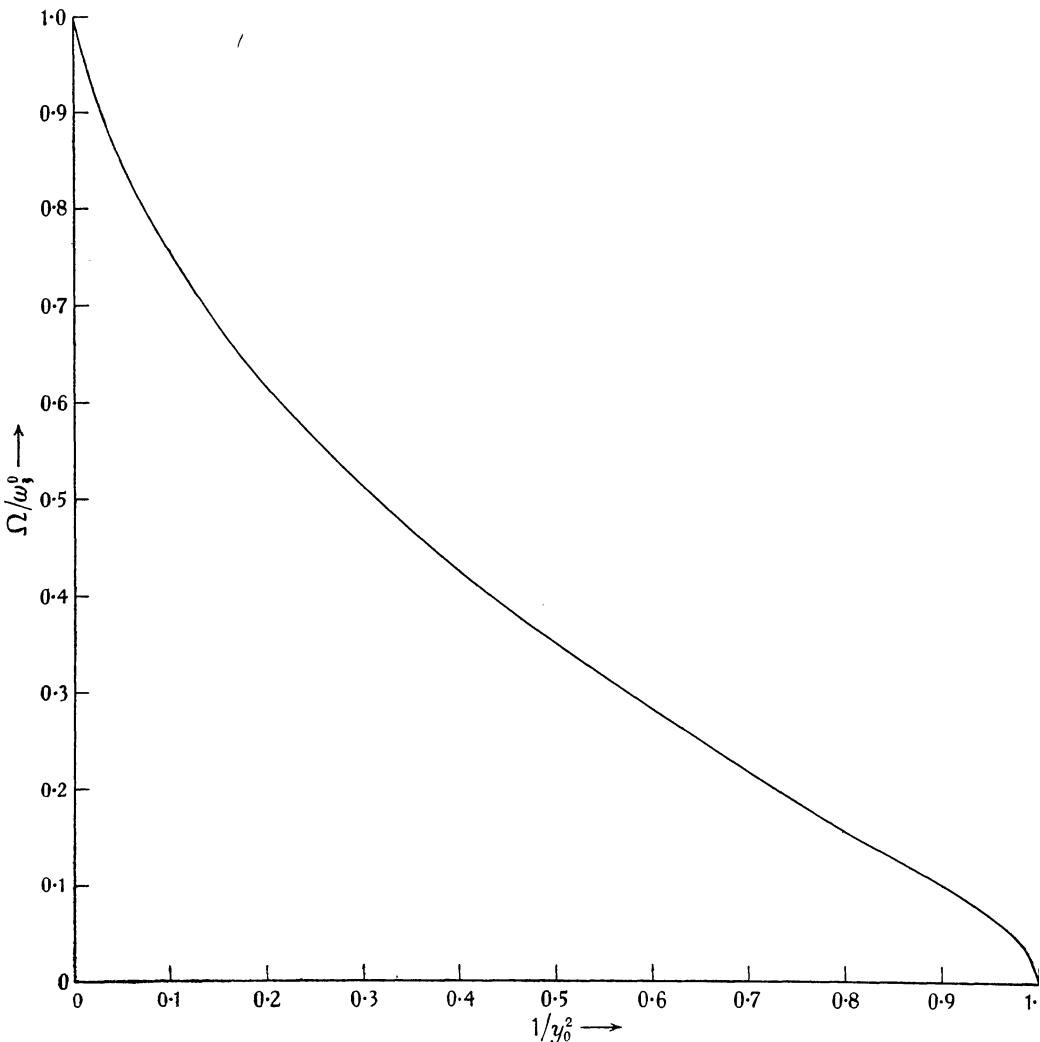


FIG. 1.— $\{\Omega/\omega_3^0, 1/y_0^2\}$ -relation.

From (77) and (77') we have on eliminating  $x_0$

$$2R_1 = \left( \frac{\omega_{3/2}^0 M_3}{\omega_3^0 M} \right)^{1/3} \cdot l_1, \quad (78)$$

where following Milne we have introduced the "invariant"  $\omega_{3/2}^0$  defined by

$$\omega_{3/2}^0 = - \left( \xi^5 \frac{d\theta_{3/2}}{d\xi} \right)_1 = 132.3843. \quad (79)$$

It is of interest to notice that the two invariants  $\omega_3^0$  and  $\omega_{3/2}^0$  of the Emden equation with the indices  $n=3$  and  $3/2$  occur in (78) in a "symmetrical way." Numerically (78) is found to be

$$R_1 = 2.01647 \left( \frac{M_3}{M} \right)^{1/3} \cdot l_1. \quad (80)$$

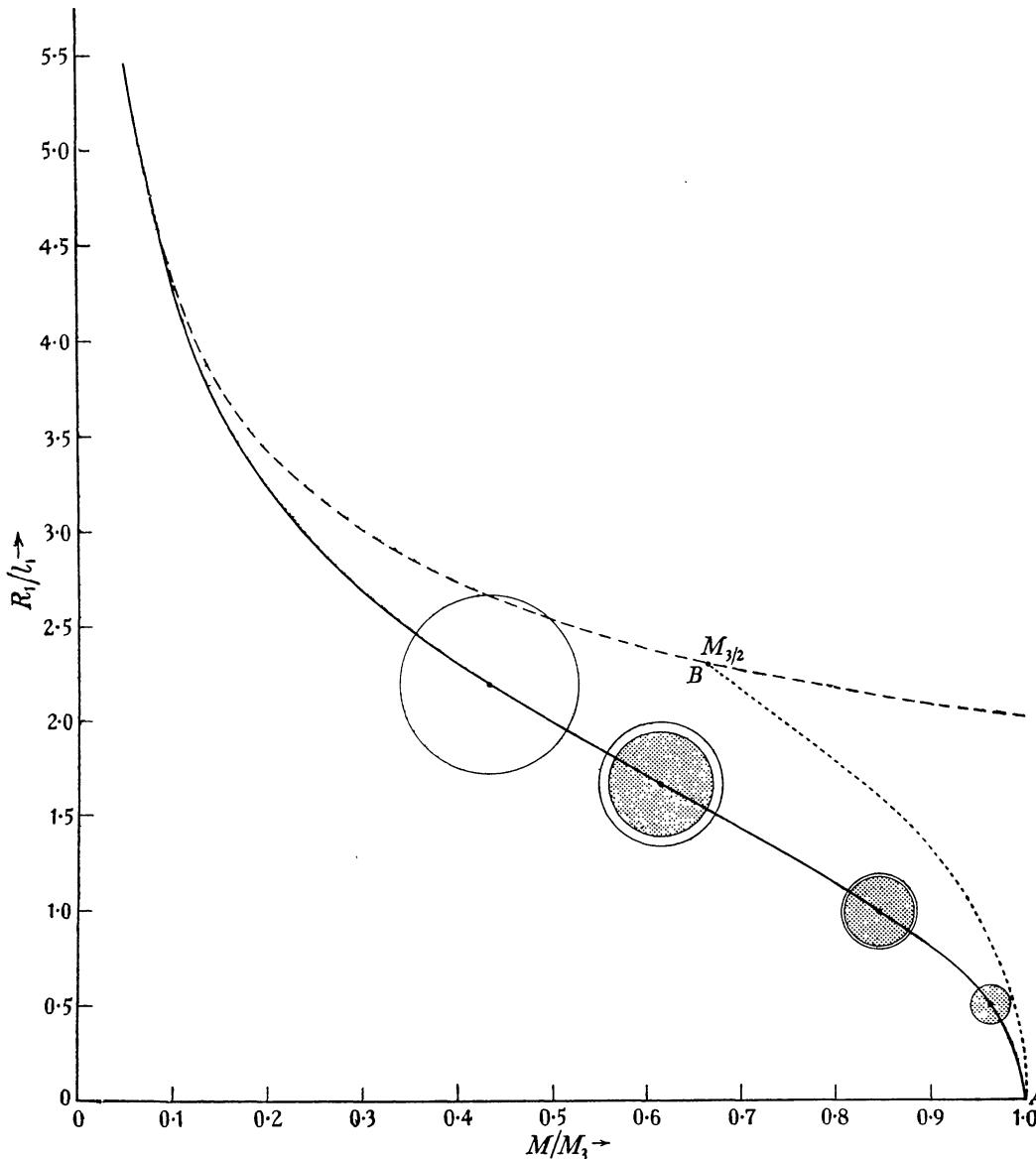


FIG. 2.—The full line curve represents the exact (mass-radius)-relation for the highly collapsed configurations. This curve tends asymptotically to the - - - curve as  $M \rightarrow 0$ .

(80) expresses the *mass-radius* relation for the polytropic limit, the radius and the mass expressed in the same units as the quantities in Table II. Similarly the mass-central density relation now reads

$$x_0^3 = 4.42381(M/M_3)^2. \quad (81)$$

The results calculated on the basis of (80) and (81) for the same masses as in Table II are summarised in Table IV. The corresponding curves are shown dotted in figs. 2 and 3.

TABLE IV

$M/M_3$	$R_1/l_1$	$x_0^3$
1	2.0165	4.4238
0.9573	2.0459	4.0538
0.9242	2.0700	3.7780
0.8471	2.1311	3.1739
0.7524	2.2174	2.5042
0.6159	2.3701	1.6778
0.5122	2.5203	1.1603
0.4260	2.6801	0.8027
0.3503	2.8606	0.5429
0.2814	3.0772	0.3502
0.1532	3.7691	0.1038

One notices clearly from these two curves how marked the deviations from the limiting curves become even for quite small masses. Thus for  $M=0.15M_3$ , the central density predicted by our exact treatment is about 25 per cent. greater and the radius about 5 per cent. smaller. The relativistic effects are therefore quite significant even for small masses. They certainly cannot be ignored for masses greater than  $0.2M_3$ . Of course the extrapolation of the  $n=3/2$  configurations for masses (in units of  $M_3$ ) approaching unity is quite misleading. These completely collapsed configurations have a natural limit, and our exact treatment now shows how this limit is reached.

It is of interest to compare the full-line curve in fig. 2 representing our exact (mass-radius) curve with what one would obtain by the methods of I, where the degenerate spheres of mass greater than a certain limit  $M_{3/2}$  were considered as "composite configurations." The mass  $M_{3/2}$  was defined as one in which the Emden polytrope with  $n=3/2$ \* would have a central density  $\rho' (=K_2/K_1)^3$ . In our present notation we have by (81)

$$M_{3/2} = \sqrt{\frac{(1.25)^3}{4.42381}} \cdot M_3 = 0.66446 M_3. \quad (82)$$

This particular point is marked as B in fig. 2 on the ---- curve. A treatment of the composite configurations by the methods of I would have led to some kind of curve like the dotted one in fig. 2 conjecturally drawn. But fortunately it is now not necessary to go into the very elaborate numerical work that would have been involved to fix the part BA by the methods of I. By a single system of integrations we have now fixed the exact nature of the (mass-radius) curve for these completely collapsed configurations.

\* The equation of state being  $p = K_1 \rho^{5/3}$ .

14. *The Relative Density Distributions in the Different Configurations.*—Our main diagram (fig. 4) now illustrates the relative density distributions in the configurations studied. Here we have plotted  $(\rho/\rho_0)$  against  $(\eta/\eta_1)$  for the different masses for which we have numerical results. The two limiting density distributions specified by Emden,  $\theta_3$  and  $\theta_{3/2}$ , are also shown (dotted) in the same figure. Fig. 4, which is the principal outcome of our studies, presents a set of ten out of a continuous family of density distributions

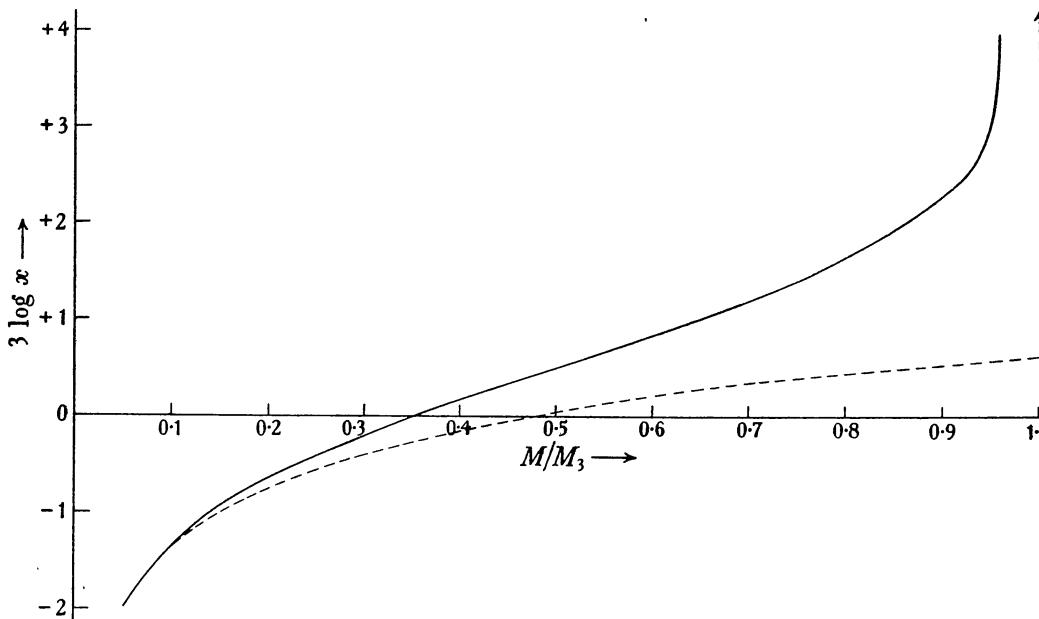


FIG. 3.—The full line curve represents the exact (mass,  $\log \rho_0$ )-relation for the highly collapsed configurations. This curve tends asymptotically to the dotted curve as  $M \rightarrow 0$ .

covering the range specified by the polytropic distributions of indices  $3/2$  and  $3$ .

15. *Concluding Remarks.*—In this paper we have strictly confined ourselves to the case “ $\beta = 1$ .” But in stellar problem the radiation pressure (even if small) necessarily plays a deciding rôle, and the question as to in what sense we have to understand the completely degenerate spheres studied here as representing “the limiting sequence of configurations to which all stars must tend eventually” can be answered only by introducing radiation in these configurations. To do this properly we have first to develop adequate methods to treat composite configurations consisting of degenerate cores (of the structures studied here) surrounded by gaseous envelopes. These and related problems are studied in the following paper (p. 226).

16. *Manuscript Copy of Tables.*—The functions  $\phi$  and their derivatives  $\phi'$  (to six and five significant figures respectively) have been computed by the author for the values of  $1/y_0^2$  specified in (73). In addition to  $\phi$  and  $\phi'$  the auxiliary functions  $\rho/\rho_0$ ,  $\rho_0/\bar{\rho}$ ,  $-\eta^2\phi'$  and two other functions  $U$  and  $V$  (defined in equation (91) of the following paper) have also been tabulated. The auxiliary functions were calculated correct to five significant figures. All

the functions were tabulated for steps of 0.1 for the argument  $\eta$ . A manuscript copy of these tables has been deposited in the Library of the Society.\*

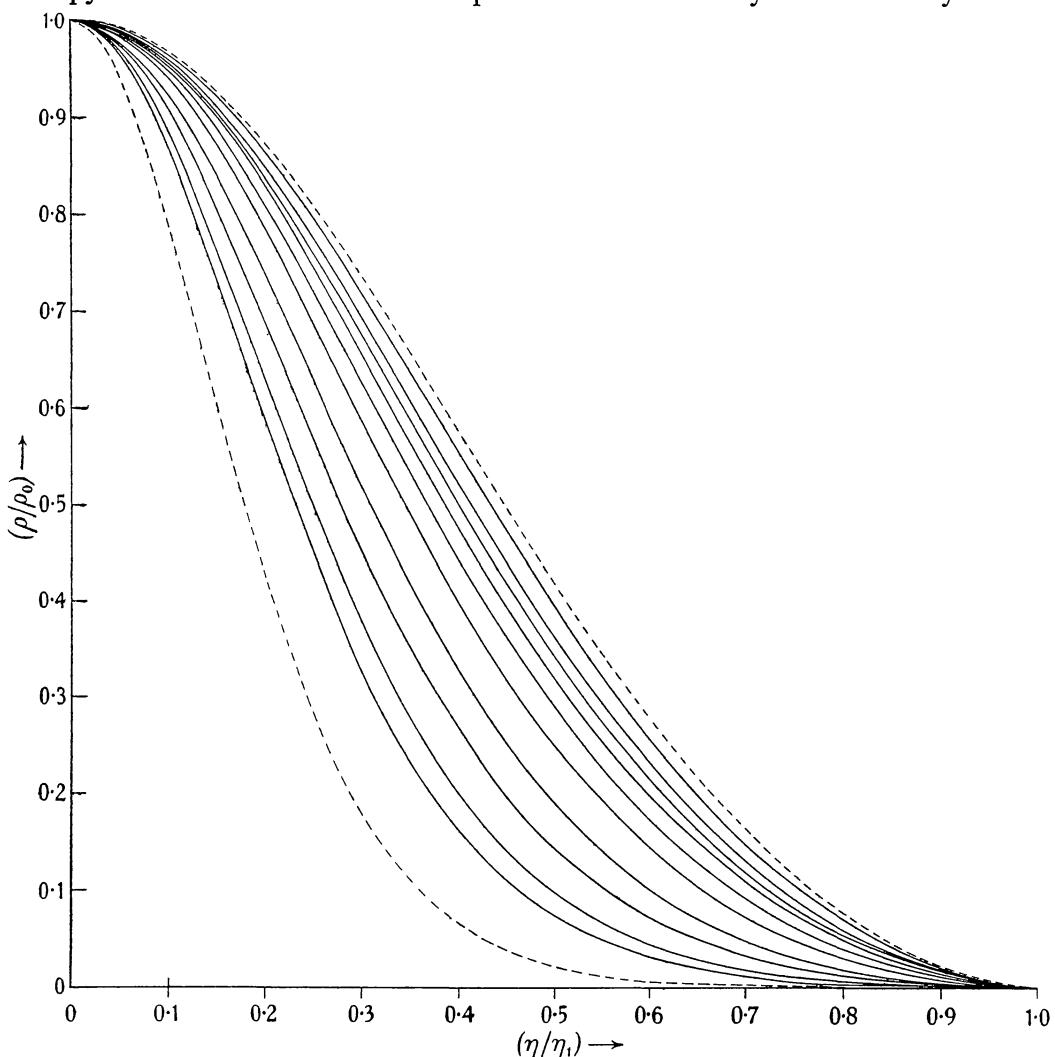


FIG. 4.—*The relative density distributions in the highly collapsed configurations. The upper dotted curve corresponds to the polytropic distribution  $n = 3/2$  and the lower dotted curve to the polytropic distribution  $n = 3$ . The inner curves represent the density distributions for  $1/y_0^2 = 0.8, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0.05, 0.02, 0.01$  respectively.*

## APPENDIX

*The Equation of State for a Degenerate Gas.*—The equation has been derived by Stoner (among others),† but we shall give a simpler derivation of the same.

In a completely degenerate electron assembly all the electrons have momenta less than a certain “threshold” value  $p_0$ , and in the region of the

\* Dr. Chandrasekhar's Tables can be consulted by Fellows on application to the Assistant Secretary (Editors).

† M.N., 92, 444, 1931.

available phase space of volume  $\frac{4}{3}\pi p_0^3 V$  every cell of volume  $h^3$  contains just two electrons. Clearly then we have

$$n = \frac{8\pi}{h^3} \int_0^{p_0} p^2 dp, \quad (1)^*$$

$$\mathfrak{E} = \frac{8\pi V}{h^3} \int_0^{p_0} E p^2 dp, \quad (2)$$

$$P = \frac{8\pi}{3h^3} \int_0^{p_0} p^3 \frac{dE}{dp} dp, \quad (3)$$

where  $n$  is the number of electrons per unit volume in the assembly of volume  $V$ ,  $\mathfrak{E}$  the total energy and  $E$  the kinetic energy of a free electron. We have now denoted the pressure by  $P$  instead of by " $p$ " as in the text of the paper to avoid confusion with the momentum, which has to be denoted by " $p$ ." From (1) and from (2) and (3) we have respectively

$$p_0^3 = \frac{3h^3 n}{8\pi}; \quad P = \frac{8\pi}{3h^3} E(p_0) p_0^3 - \frac{\mathfrak{E}}{V}. \quad (4)$$

Equations (1) to (4) are quite general. Now in the relativistic mechanics we have

$$E = mc^2 \left\{ \left( 1 + \frac{p^2}{m^2 c^2} \right)^{1/2} - 1 \right\}, \quad (5)$$

or

$$p^2 = \frac{E(E + 2mc^2)}{c^2}. \quad (5')$$

Using (5') in (3) we have, after some minor transformations, that

$$P = \frac{8\pi m^4 c^5}{3h^3} \int_0^{\theta_0} \sinh^4 \theta d\theta, \quad (6)$$

where

$$\sinh \theta = p/mc; \quad \sinh \theta_0 = p_0/mc. \quad (7)\dagger$$

(7) yields at once that

$$P = \frac{8\pi m^4 c^5}{3h^3} \left[ \frac{\sinh^3 \theta \cosh \theta}{4} - \frac{3}{16} \sinh 2\theta + \frac{3}{8} \theta \right]_{\theta=\theta_0}. \quad (8)$$

Writing  $x$  for  $(p_0/mc)$  we have

$$P = \frac{\pi m^4 c^5}{3h^3} \left[ x(2x^2 - 3)(x^2 + 1)^{1/2} + 3 \sinh^{-1} x \right], \quad (9)$$

\* This equation follows directly from the expression for the number of waves associated with electrons whose energies lie between  $E$  and  $E + dE$  given by Dirac (*P.R.S.*, 112, 660, 1926, his unnumbered equation on p. 671). Actually Dirac obtains this result using the Klein-Gordon relativistic wave equation. That the same result would follow from Dirac's relativistic wave equation (on neglecting the states of kinetic energy—which is permissible when no external perturbations are present) is clear from J. von Neumann, *Z. f. Physik*, 48, 868, 1928.

†  $\theta$  here introduced will not be confused with the Emden function.

$$\rho = n\mu H = \frac{8\pi m^3 c^3 \mu H}{3h^3} x^3, \quad (10)$$

which are the equations quoted in the text. Our derivation now shows "why" we are able to reduce the differential equation for degenerate gas spheres in gravitational equilibrium to such a simple form. The "reason" is that we have such an elementary integral for  $P$  as in (6).

The function  $f(x)$  on the right-hand side of (9) has the following asymptotic forms :—

$$f(x) \sim \frac{8}{5}x^5 - \frac{4}{7}x^7 + \frac{1}{3}x^9 - \frac{5}{2}x^{11} + \dots \quad x \rightarrow 0, \quad (11)$$

$$f(x) \sim 2x^4 - 3x^2 + \dots \quad x \rightarrow \infty. \quad (12)$$

Finally we notice that

$$\frac{f(x)}{2x^4} < 1 \quad \text{for all finite } x. \quad (13)$$

The inequality in (13) is a *strict* one. If only the first terms in the expansions (11) and (12) are retained, we can easily eliminate  $x$  from (9) and (10) for these limiting cases and obtain, as we should expect, that

$$P = K_1 \rho^{5/3} \quad (x \rightarrow 0); \quad P = K_2 \rho^{4/3} \quad (x \rightarrow \infty), \quad (14)$$

with

$$K_1 = \frac{1}{20} \left( \frac{3}{\pi} \right)^{2/3} \frac{h^2}{m(\mu H)^{5/3}}; \quad K_2 = \left( \frac{3}{\pi} \right)^{1/3} \frac{hc}{8(\mu H)^{4/3}}. \quad (15)^*$$

If we write our "equation of state" (9) and (10) parametrically as (changing to "p" to denote pressure),

$$p = A_2 f(x); \quad \rho = B x^3, \quad (16)$$

we find, on putting in the numerical values for the constants, that (in C.G.S. units)

$$A_2 = 6.0406 \times 10^{22}; \quad B = 9.8848 \times 10^5 \mu, \quad (17)$$

or

$$\begin{aligned} \log \rho &= 5.9950 + 3 \log x + \log \mu, \\ \log p &= 22.7811 + \log f(x) \end{aligned} \quad (18)$$

Stoner has previously made some calculations concerning the  $(p, \rho)$  relation for a degenerate gas, but for the study in the following paper more accurate tables for  $f(x)$  were needed. Accordingly the whole computation was re-

\* The law  $P = K_2 \rho^{4/3}$  was first used by the author in his paper on "Highly Collapsed Configurations," etc. (*M.N.*, 91, 456, 1931). This law has also been derived by E. C. Stoner (*M.N.*, 92, 444, 1932), T. E. Sterne (*M.N.*, 93, 764, 1933), and is also implicitly contained in J. Frenkel (*Z. f. Physik*, 50, 234, 1928). The law has also been used by L. Landau (*Physik. Zeits. d. Soviet Union*, 1, 285, 1932). It may also be pointed out that the law  $P = K_2 \rho^{4/3}$  is implicit in certain equations in a paper by F. Juttner (*Z. f. Physik*, 47, 542, 1928, equations in §§ 13, 17; our equation (6) above is a limiting form of Juttner's integral  $Q(a, \gamma; +1)$ ). This last work of Juttner is related to his earlier work on the relativistic theory of an ideal classical gas, for a convenient summary of which see W. Pauli, *Relativitätstheorie* (Leipzig, Teubner), § 49.

done and the results are tabulated in Table V. I am indebted to Dr. Comrie and Mr. Sadler for the loan of a manuscript copy of a seven-figure table for  $\sinh^{-1} x$ , which was valuable in the computations of  $f(x)$ .

TABLE V

$x$	$f(x)$	$f(x)/2x^4$
0	0	0
0.2	0.000505	0.15785
0.4	0.015527	.39325
0.6	0.111126	.42873
0.8	0.435865	.53206
1.0	1.229907	.61495
1.2	2.82298	.68070
1.4	5.62991	.73276
1.6	10.14696	.77415
1.8	16.94969	.80731
2.0	26.69159	.83411
2.2	40.10347	.85598
2.4	57.99311	.87398
2.6	81.24509	.88894
2.8	110.8207	.90149
3.0	147.7578	.91209
3.5	279.8113	.93232
4.0	484.5644	.94641
4.5	784.5271	.95659
5.0	1205.2069	.96417
6.0	2525.739	.97444
7.0	4710.192	.98088
8.0	8070.587	.98518
9.0	$1.296694 \times 10^4$	.98818
10.0	$1.980725 \times 10^4$	.99036
20.0	$3.192093 \times 10^5$	.99753
30.0	$1.618212 \times 10^6$	.99890
40.0	$5.116812 \times 10^6$	.99938
50.0	$1.249501 \times 10^7$	.99960
60.0	$2.591280 \times 10^7$	.99972
70.0	$4.801018 \times 10^7$	.99980
80.0	$8.190727 \times 10^7$	.99984
90.0	$13.12039 \times 10^7$	.99988
100.0	$19.9980 \times 10^7$	.99990

Trinity College, Cambridge :  
1935 January 1.



MAY 15, 1935

PHYSICAL REVIEW

VOLUME 47

## Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?

A. EINSTEIN, B. PODOLSKY AND N. ROSEN, *Institute for Advanced Study, Princeton, New Jersey*

(Received March 25, 1935)

In a complete theory there is an element corresponding to each element of reality. A sufficient condition for the reality of a physical quantity is the possibility of predicting it with certainty, without disturbing the system. In quantum mechanics in the case of two physical quantities described by non-commuting operators, the knowledge of one precludes the knowledge of the other. Then either (1) the description of reality given by the wave function in

quantum mechanics is not complete or (2) these two quantities cannot have simultaneous reality. Consideration of the problem of making predictions concerning a system on the basis of measurements made on another system that had previously interacted with it leads to the result that if (1) is false then (2) is also false. One is thus led to conclude that the description of reality as given by a wave function is not complete.

### 1.

**A**NY serious consideration of a physical theory must take into account the distinction between the objective reality, which is independent of any theory, and the physical concepts with which the theory operates. These concepts are intended to correspond with the objective reality, and by means of these concepts we picture this reality to ourselves.

In attempting to judge the success of a physical theory, we may ask ourselves two questions: (1) "Is the theory correct?" and (2) "Is the description given by the theory complete?" It is only in the case in which positive answers may be given to both of these questions, that the concepts of the theory may be said to be satisfactory. The correctness of the theory is judged by the degree of agreement between the conclusions of the theory and human experience. This experience, which alone enables us to make inferences about reality, in physics takes the form of experiment and measurement. It is the second question that we wish to consider here, as applied to quantum mechanics.

Whatever the meaning assigned to the term *complete*, the following requirement for a complete theory seems to be a necessary one: *every element of the physical reality must have a counterpart in the physical theory*. We shall call this the condition of completeness. The second question is thus easily answered, as soon as we are able to decide what are the elements of the physical reality.

The elements of the physical reality cannot be determined by *a priori* philosophical considerations, but must be found by an appeal to results of experiments and measurements. A comprehensive definition of reality is, however, unnecessary for our purpose. We shall be satisfied with the following criterion, which we regard as reasonable. *If, without in any way disturbing a system, we can predict with certainty (i.e., with probability equal to unity) the value of a physical quantity, then there exists an element of physical reality corresponding to this physical quantity.* It seems to us that this criterion, while far from exhausting all possible ways of recognizing a physical reality, at least provides us with one

such way, whenever the conditions set down in it occur. Regarded not as a necessary, but merely as a sufficient, condition of reality, this criterion is in agreement with classical as well as quantum-mechanical ideas of reality.

To illustrate the ideas involved let us consider the quantum-mechanical description of the behavior of a particle having a single degree of freedom. The fundamental concept of the theory is the concept of *state*, which is supposed to be completely characterized by the wave function  $\psi$ , which is a function of the variables chosen to describe the particle's behavior. Corresponding to each physically observable quantity  $A$  there is an operator, which may be designated by the same letter.

If  $\psi$  is an eigenfunction of the operator  $A$ , that is, if

$$\psi' \equiv A\psi = a\psi, \quad (1)$$

where  $a$  is a number, then the physical quantity  $A$  has with certainty the value  $a$  whenever the particle is in the state given by  $\psi$ . In accordance with our criterion of reality, for a particle in the state given by  $\psi$  for which Eq. (1) holds, there is an element of physical reality corresponding to the physical quantity  $A$ . Let, for example,

$$\psi = e^{(2\pi i/h)p_0x}, \quad (2)$$

where  $h$  is Planck's constant,  $p_0$  is some constant number, and  $x$  the independent variable. Since the operator corresponding to the momentum of the particle is

$$p = (h/2\pi i)\partial/\partial x, \quad (3)$$

we obtain

$$\psi' = p\psi = (h/2\pi i)\partial\psi/\partial x = p_0\psi. \quad (4)$$

Thus, in the state given by Eq. (2), the momentum has certainly the value  $p_0$ . It thus has meaning to say that the momentum of the particle in the state given by Eq. (2) is real.

On the other hand if Eq. (1) does not hold, we can no longer speak of the physical quantity  $A$  having a particular value. This is the case, for example, with the coordinate of the particle. The operator corresponding to it, say  $q$ , is the operator of multiplication by the independent variable. Thus,

$$q\psi = x\psi \neq a\psi. \quad (5)$$

In accordance with quantum mechanics we can only say that the relative probability that a measurement of the coordinate will give a result lying between  $a$  and  $b$  is

$$P(a, b) = \int_a^b |\psi|^2 dx = \int_a^b dx = b - a. \quad (6)$$

Since this probability is independent of  $a$ , but depends only upon the difference  $b - a$ , we see that all values of the coordinate are equally probable.

A definite value of the coordinate, for a particle in the state given by Eq. (2), is thus not predictable, but may be obtained only by a direct measurement. Such a measurement however disturbs the particle and thus alters its state. After the coordinate is determined, the particle will no longer be in the state given by Eq. (2). The usual conclusion from this in quantum mechanics is that *when the momentum of a particle is known, its coordinate has no physical reality*.

More generally, it is shown in quantum mechanics that, if the operators corresponding to two physical quantities, say  $A$  and  $B$ , do not commute, that is, if  $AB \neq BA$ , then the precise knowledge of one of them precludes such a knowledge of the other. Furthermore, any attempt to determine the latter experimentally will alter the state of the system in such a way as to destroy the knowledge of the first.

From this follows that either (1) *the quantum-mechanical description of reality given by the wave function is not complete* or (2) *when the operators corresponding to two physical quantities do not commute the two quantities cannot have simultaneous reality*. For if both of them had simultaneous reality—and thus definite values—these values would enter into the complete description, according to the condition of completeness. If then the wave function provided such a complete description of reality, it would contain these values; these would then be predictable. This not being the case, we are left with the alternatives stated.

In quantum mechanics it is usually assumed that the wave function *does* contain a complete description of the physical reality of the system in the state to which it corresponds. At first

sight this assumption is entirely reasonable, for the information obtainable from a wave function seems to correspond exactly to what can be measured without altering the state of the system. We shall show, however, that this assumption, together with the criterion of reality given above, leads to a contradiction.

## 2.

For this purpose let us suppose that we have two systems, I and II, which we permit to interact from the time  $t=0$  to  $t=T$ , after which time we suppose that there is no longer any interaction between the two parts. We suppose further that the states of the two systems before  $t=0$  were known. We can then calculate with the help of Schrödinger's equation the state of the combined system I+II at any subsequent time; in particular, for any  $t>T$ . Let us designate the corresponding wave function by  $\Psi$ . We cannot, however, calculate the state in which either one of the two systems is left after the interaction. This, according to quantum mechanics, can be done only with the help of further measurements, by a process known as the *reduction of the wave packet*. Let us consider the essentials of this process.

Let  $a_1, a_2, a_3, \dots$  be the eigenvalues of some physical quantity  $A$  pertaining to system I and  $u_1(x_1), u_2(x_1), u_3(x_1), \dots$  the corresponding eigenfunctions, where  $x_1$  stands for the variables used to describe the first system. Then  $\Psi$ , considered as a function of  $x_1$ , can be expressed as

$$\Psi(x_1, x_2) = \sum_{n=1}^{\infty} \psi_n(x_2) u_n(x_1), \quad (7)$$

where  $x_2$  stands for the variables used to describe the second system. Here  $\psi_n(x_2)$  are to be regarded merely as the coefficients of the expansion of  $\Psi$  into a series of orthogonal functions  $u_n(x_1)$ . Suppose now that the quantity  $A$  is measured and it is found that it has the value  $a_k$ . It is then concluded that after the measurement the first system is left in the state given by the wave function  $u_k(x_1)$ , and that the second system is left in the state given by the wave function  $\psi_k(x_2)$ . This is the process of reduction of the wave packet; the wave packet given by the

infinite series (7) is reduced to a single term  $\psi_k(x_2)u_k(x_1)$ .

The set of functions  $u_n(x_1)$  is determined by the choice of the physical quantity  $A$ . If, instead of this, we had chosen another quantity, say  $B$ , having the eigenvalues  $b_1, b_2, b_3, \dots$  and eigenfunctions  $v_1(x_1), v_2(x_1), v_3(x_1), \dots$  we should have obtained, instead of Eq. (7), the expansion

$$\Psi(x_1, x_2) = \sum_{s=1}^{\infty} \varphi_s(x_2) v_s(x_1), \quad (8)$$

where  $\varphi_s$ 's are the new coefficients. If now the quantity  $B$  is measured and is found to have the value  $b_r$ , we conclude that after the measurement the first system is left in the state given by  $v_r(x_1)$  and the second system is left in the state given by  $\varphi_r(x_2)$ .

We see therefore that, as a consequence of two different measurements performed upon the first system, the second system may be left in states with two different wave functions. On the other hand, since at the time of measurement the two systems no longer interact, no real change can take place in the second system in consequence of anything that may be done to the first system. This is, of course, merely a statement of what is meant by the absence of an interaction between the two systems. Thus, *it is possible to assign two different wave functions* (in our example  $\psi_k$  and  $\varphi_r$ ) *to the same reality* (the second system after the interaction with the first).

Now, it may happen that the two wave functions,  $\psi_k$  and  $\varphi_r$ , are eigenfunctions of two non-commuting operators corresponding to some physical quantities  $P$  and  $Q$ , respectively. That this may actually be the case can best be shown by an example. Let us suppose that the two systems are two particles, and that

$$\Psi(x_1, x_2) = \int_{-\infty}^{\infty} e^{(2\pi i/\hbar)(x_1 - x_2 + x_0)p} dp, \quad (9)$$

where  $x_0$  is some constant. Let  $A$  be the momentum of the first particle; then, as we have seen in Eq. (4), its eigenfunctions will be

$$u_p(x_1) = e^{(2\pi i/\hbar)px_1} \quad (10)$$

corresponding to the eigenvalue  $p$ . Since we have here the case of a continuous spectrum, Eq. (7) will now be written

$$\Psi(x_1, x_2) = \int_{-\infty}^{\infty} \psi_p(x_2) u_p(x_1) dp, \quad (11)$$

where

$$\psi_p(x_2) = e^{-(2\pi i/\hbar)(x_2 - x_0)p}. \quad (12)$$

This  $\psi_p$ , however, is the eigenfunction of the operator

$$P = (\hbar/2\pi i) \partial/\partial x_2, \quad (13)$$

corresponding to the eigenvalue  $-p$  of the momentum of the second particle. On the other hand, if  $B$  is the coordinate of the first particle, it has for eigenfunctions

$$v_x(x_1) = \delta(x_1 - x), \quad (14)$$

corresponding to the eigenvalue  $x$ , where  $\delta(x_1 - x)$  is the well-known Dirac delta-function. Eq. (8) in this case becomes

$$\Psi(x_1, x_2) = \int_{-\infty}^{\infty} \varphi_x(x_2) v_x(x_1) dx, \quad (15)$$

where

$$\begin{aligned} \varphi_x(x_2) &= \int_{-\infty}^{\infty} e^{(2\pi i/\hbar)(x - x_2 + x_0)p} dp \\ &= \hbar \delta(x - x_2 + x_0). \end{aligned} \quad (16)$$

This  $\varphi_x$ , however, is the eigenfunction of the operator

$$Q = x_2 \quad (17)$$

corresponding to the eigenvalue  $x + x_0$  of the coordinate of the second particle. Since

$$PQ - QP = \hbar/2\pi i, \quad (18)$$

we have shown that it is in general possible for  $\psi_k$  and  $\varphi_r$  to be eigenfunctions of two noncommuting operators, corresponding to physical quantities.

Returning now to the general case contemplated in Eqs. (7) and (8), we assume that  $\psi_k$  and  $\varphi_r$  are indeed eigenfunctions of some noncommuting operators  $P$  and  $Q$ , corresponding to the eigenvalues  $p_k$  and  $q_r$ , respectively. Thus, by measuring either  $A$  or  $B$  we are in a position to predict with certainty, and without in any way

disturbing the second system, either the value of the quantity  $P$  (that is  $p_k$ ) or the value of the quantity  $Q$  (that is  $q_r$ ). In accordance with our criterion of reality, in the first case we must consider the quantity  $P$  as being an element of reality, in the second case the quantity  $Q$  is an element of reality. But, as we have seen, both wave functions  $\psi_k$  and  $\varphi_r$  belong to the same reality.

Previously we proved that either (1) the quantum-mechanical description of reality given by the wave function is not complete or (2) when the operators corresponding to two physical quantities do not commute the two quantities cannot have simultaneous reality. Starting then with the assumption that the wave function does give a complete description of the physical reality, we arrived at the conclusion that two physical quantities, with noncommuting operators, can have simultaneous reality. Thus the negation of (1) leads to the negation of the only other alternative (2). We are thus forced to conclude that the quantum-mechanical description of physical reality given by wave functions is not complete.

One could object to this conclusion on the grounds that our criterion of reality is not sufficiently restrictive. Indeed, one would not arrive at our conclusion if one insisted that two or more physical quantities can be regarded as simultaneous elements of reality *only when they can be simultaneously measured or predicted*. On this point of view, since either one or the other, but not both simultaneously, of the quantities  $P$  and  $Q$  can be predicted, they are not simultaneously real. This makes the reality of  $P$  and  $Q$  depend upon the process of measurement carried out on the first system, which does not disturb the second system in any way. No reasonable definition of reality could be expected to permit this.

While we have thus shown that the wave function does not provide a complete description of the physical reality, we left open the question of whether or not such a description exists. We believe, however, that such a theory is possible.

## On the Interaction of Elementary Particles

H. Yukawa

(Received 1935)

At the present stage of the quantum theory little is known about the nature of interaction of elementary particles, Heisenberg considered the interaction of "Platzwechsel" between the neutron and the proton to be of importance to the nuclear structure.

Recently Fermi treated the problem of  $\beta$ -disintegration on the hypothesis of "neutrino". According to this theory, the neutron and the proton can interact by emitting and absorbing a pair of neutrino and electron. Unfortunately the interaction energy calculated on such assumption is much too small to account for the binding energies of neutrons and protons in the nucleus.

To remove this defect, it seems natural to modify the theory of Heisenberg and Fermi in the following way. The transition of a heavy particle from neutron state to proton state is not always accompanied by the emission of light particles, i.e., a neutrino and an electron, but the energy liberated by the transition is taken up sometimes by another heavy particle, which in turn will be transformed from proton state into neutron state. If the probability of occurrence of the latter process is much larger than that of the former, the interaction between the neutron and the proton will be much larger than in the case of Fermi, whereas the probability of emission of light particles is not affected essentially.

Now such interaction between the elementary particles can be described by means of a field of force, just as the interaction between the charged particles is described by the electromagnetic field. The above considerations show that the interaction of heavy particles with this field is much larger than that of light particles with it.

In the quantum theory this field should be accompanied by a new sort of quantum, just as the electromagnetic field is accompanied by the photon.

In this paper the possible natures of this field and the quantum accompanying it will be discussed briefly and also their bearing on the nuclear structure will be considered.

Besides such an exchange force and the ordinary electric and magnetic forces there may be other forces between the elementary particles, but we disregard the latter for the moment.

Fuller account will be made in the next paper.

## Field Describing the Interaction

In analogy with the scalar potential of the electromagnetic field, a function  $U(x, y, z, t)$  is introduced to describe the field between the neutron and the proton. This function will satisfy an equation similar to the wave equation for the electromagnetic potential.

Now the equation

$$\left\{ \Delta - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \right\} U = 0 \quad (1)$$

has only static solution with central symmetry  $\frac{1}{r}$ , except the additive and the multiplicative constants. The potential of force between the neutron and proton should, however, not be of Coulomb type, but decrease more rapidly with distance. It can be expressed, for example by

$$+ \text{or} - g^2 \frac{e^{-\lambda r}}{r}, \quad (2)$$

where  $g$  is a constant with the dimension of electric charge, i.e.,  $\text{cm.}^{3/2} \text{ sec.}^{-1} \text{ gr.}^{1/2}$  and  $\lambda$  with the dimension  $\text{cm.}^{-1}$

Since this function is a static with central symmetry of the wave equation

$$\left\{ \Delta - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} - \lambda^2 \right\} U = 0, \quad (3)$$

let this equation be assumed to be the correct equation for  $U$  in vacuum. In the presence of the heavy particles, the  $U$ -field interacts with them and causes the transition from neutron state to proton state.

Now, if we introduce the matrices

$$\tau_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \tau_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \tau_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

and denote the neutron state and the proton state by  $\tau_3 = 1$  and  $\tau_3 = -1$  respectively, the wave equation is given by

$$\left\{ \Delta - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} - \lambda^2 \right\} U = -4\pi g \tilde{\Psi} \frac{\tau_1 - i\tau_2}{2} \Psi, \quad (4)$$

where  $\Psi$  denoted the wave function of the heavy particles, being a function of time, position, spin as well as  $\tau'_3$ , which takes the value either 1 or -1.

Next, the conjugate complex function  $\tilde{U}(x, y, z, t)$ , satisfying the equation

$$\left\{ \Delta - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} - \lambda^2 \right\} \tilde{U} = -4\pi g \tilde{\Psi} \frac{\tau_1 + i\tau_2}{2} \Psi, \quad (5)$$

is introduced, corresponding to the inverse transition from proton to neutron state.

Similar equation will hold for the vector function, which is the analogue of the vector potential of the electromagnetic field. However, we disregard it for the moment, as there's no correct relativistic theory for the heavy particles. Hence simple non-relativistic wave equation neglecting spin will be used for the heavy particle, it the following way

$$\left\{ \frac{h^2}{4} \left( \frac{1+\tau_3}{M_N} + \frac{1-\tau_3}{M_P} \right) \Delta + ih \frac{\partial}{\partial t} - \frac{1+\tau_3}{2} M_N c^2 - \frac{1-\tau_3}{2} M_P c^2 - g \left( \tilde{U} \frac{\tau_1 - i\tau_2}{2} + U \frac{\tau_1 + i\tau_2}{2} \right) \right\} \Psi = 0, \quad (6)$$

where  $h$  is Planck's constant divided by  $2\pi$  and  $M_N, M_P$  are the masses of the neutron and the proton respectively. The reason for taking the negative sign in front of  $g$  will be mentioned later.

The equation (6) corresponds to the Hamiltonian

$$H = \left( \frac{1+\tau_3}{4M_N} + \frac{1-\tau_3}{4M_P} \right) \vec{p}^2 + \frac{1+\tau_3}{2} M_N c^2 + \frac{1-\tau_3}{2} M_P c^2 + g \left( \tilde{U} \frac{\tau_1 - i\tau_2}{2} + U \frac{\tau_1 + i\tau_2}{2} \right) \quad (7)$$

where  $\vec{p}$  is the momentum of the particle. If we put  $M_N c^2 - M_P c^2 = D$  and  $M_N + M_P = 2M$ , the equation (7) becomes approximately

$$H = \frac{\vec{p}^2}{2M} + \frac{g}{2} \left\{ \tilde{U} (\tau_1 - i\tau_2) + U (\tau_1 + i\tau_2) \right\} + \frac{D}{2} \tau_3, \quad (8)$$

where the constant term  $M c^2$  omitted.

Now consider two heavy particles at point  $(x_1, y_1, z_1)$  and  $(x_2, y_2, z_2)$  respectively and assume their relative velocity to be small. The field at  $(x_1, y_1, z_1)$  due to the particle at  $(x_2, y_2, z_2)$  are, from (4) and (5),

$$\left. \begin{aligned} U(x_1, y_1, z_1) &= g \frac{e^{-\lambda\tau_{12}}}{\tau_{12}} \frac{(\tau_1^{(2)} - i\tau_2^{(2)})}{2} \\ \text{and} \\ \tilde{U}(x_1, y_1, z_1) &= g \frac{e^{-\lambda\tau_{12}}}{\tau_{12}} \frac{(\tau_1^{(2)} - i\tau_2^{(2)})}{2}, \end{aligned} \right\} \quad (9)$$

where  $(\tau_1^{(1)}, \tau_2^{(1)}, \tau_3^{(1)})$  and  $(\tau_1^{(2)}, \tau_2^{(2)}, \tau_3^{(2)})$  are the matrices relating to the first and the second particles respectively, and  $\tau_{12}$  is the distance between them.

Hence the Hamiltonian for the system is given, in the absence of the external fields by,

$$H = \frac{\vec{p}_1^2}{2M} + \frac{\vec{p}_2^2}{2M} + \frac{g^2}{4} \left\{ \left( \tau_1^{(1)} - i\tau_2^{(1)} \right) \left( \tau_1^{(2)} + i\tau_2^{(2)} \right) + \left( \tau_1^{(1)} + i\tau_2^{(1)} \right) \left( \tau_1^{(2)} - i\tau_2^{(2)} \right) \right\} \frac{e^{-\lambda\tau_{12}}}{\tau_{12}} + \left( \tau_3^{(1)} + \tau_3^{(2)} \right) D = \frac{\vec{p}_1^2}{2M} + \frac{\vec{p}_2^2}{2M} + \frac{g^2}{2} \left( \tau_1^{(1)} \tau_1^{(2)} + \tau_2^{(1)} \tau_2^{(2)} \right) \frac{e^{-\lambda\tau_{12}}}{\tau_{12}} + \left( \tau_3^{(1)} + \tau_3^{(2)} \right) D, \quad (10)$$

where  $\vec{p}_1, \vec{p}_2$  are the momenta of the particles.

This Hamiltonian is equivalent to Heisenberg's Hamiltonian, if we take for "Platzwechselintegral"

$$J(\tau) = -g^2 \frac{e^{-\lambda r}}{r}, \quad (11)$$

except that the interaction between the neutrons and the electrostatic repulsion between the protons are not taken into account. Heisenberg took the positive sign for  $J(r)$ , so that the spin of the lowest energy state of  $H^2$  was 0, whereas in our case, owing to the negative sign in front of  $g^2$ , the lowest energy state has the spin 1, which is required from the experiment.

Two constants  $g$  and  $\lambda$  appearing in the above equations should be determined by comparison with experiment. For example, using the Hamiltonian (10) for heavy particles, we can calculate the mass defect of  $H^2$  and the probability of scattering of a neutron by a proton provided that the relative velocity is small compared with the light velocity.

Rough estimation shows that the calculated values agree with the experimental results, if we take for  $\lambda$  the value between  $10^{12}$  cm $^{-1}$ . and  $10^{13}$  cm $^{-1}$ . and for  $g$  a few times of the elementary charge  $e$ , although no direct relation between  $g$  and  $e$  was suggested in the above considerations.

## Nature of the Quanta Accompanying the Field

The  $U$ -field above considered should be quantized according to the general method of the quantum theory. since the neutron and the proton both obey fermi's statistics, the quanta accompanying the  $U$ -field should obey Bose's statistics and the quantization can be carried out the line similar to that of the electromagnetic field.

The law of conservation of the electric charge demands that the quantum should have charge either  $+e$  or  $-e$ . The field quantity  $U$  corresponds to the operator which increases the number of negatively charged quanta and decreases the number of positively charged quanta by one respectively.  $\tilde{U}$ , which is the complex conjugate of  $U$ , corresponds to the inverse operator.

Next, denoting

$$p_x = -ih \frac{\partial}{\partial x}, \quad \text{etc.}, \quad W = ih \frac{\partial}{\partial t},$$

$$m_U c = \lambda h,$$

the wave equation for  $U$  in free space can be written in the form

$$\left\{ p_x^2 + p_y^2 + p_z^2 - \frac{W^2}{c^2} + m_U c^2 \right\} U = 0, \quad (12)$$

so that the quantum accompanying the field has the proper mass  $m_U = \frac{\lambda h}{c}$ .

Assuming  $\lambda = 5 \times 10^{12} \text{ cm}^{-1}$ , we obtain for  $m_U$  a value  $2 \times 10^2$  times as large as the electron mass. As such a quantum with large mass and positive or negative charge has never been found by the experiment, the above theory seems to be on a wrong line. We can show, however, that, in the ordinary nuclear transformation, such a quantum can not be emitted into outer space.

Let us consider, for example, the transition from a neutron state of energy  $W_N$  to a proton state of energy  $W_P$ , both of which include the proper energies. These states can be expressed by the wave function

$$\Psi_N(x, y, z, t, 1) = u(x, y, z) e^{-iW_N t/h}, \quad \Psi_N(x, y, z, t, -1) = 0$$

and

$$\Psi_P(x, y, z, t, 1) = 0, \quad \Psi_P(x, y, z, t, -1) = v(x, y, z) e^{-iW_P t/h},$$

so that, on the right hand side of the equation (4), the term

$$-4\pi g \tilde{\nu} u e^{-it(W_N - W_P)/h}$$

appears.

Putting  $U = U'(x, y, z) e^{i\omega t}$ , we have from (4)

$$\left\{ \Delta - \left( \lambda^2 - \frac{\omega^2}{c^2} \right) \right\} U' = -4\pi g \tilde{\nu} u, \quad (13)$$

where  $\omega = \frac{W_N - W_P}{h}$ . Integrating this, we obtain a solution

$$U'(\vec{r}) = g \int \int \int \frac{e^{-\mu|r-r'|}}{| \vec{r} - \vec{r}' |} \tilde{\nu}(\vec{r}') u(\vec{r}') d\nu', \quad (14)$$

$$\text{where } \mu = \sqrt{\lambda^2 - \frac{\omega^2}{c^2}}.$$

If  $\lambda > \frac{|\omega|}{c}$  or  $m_U c^2 > |W_N - W_P|$ ,  $\mu$  is real and the function  $J(r)$  of Heisenberg has the form  $-g^2 \frac{e^{-\mu r}}{r}$ , in which  $\mu$ , however, depends on  $|W_N - W_P|$ , becoming smaller and smaller as the latter approaches  $m_U c^2$ . This means that the range of interaction between a neutron and a proton increases as  $|W_N - W_P|$  increases.

Now the scattering (elastic or inelastic) of a neutron by a nucleus can be considered as the result of the following double process: the neutron falls into a proton level in the nucleus and a proton in the latter jumps to a neutron state of positive kinetic energy, the total energy being conserved throughout the process. The above argument, then shows that the probability of scattering may in some cases increase with the velocity of the neutron.

According to the experiment of Bonner, the collision cross section of the neutron increases, in fact, with the velocity in the case of lead whereas it decreases in the case of carbon and hydrogen, the rate of decrease being slower in the former than the latter. The origin of this effect is not clear, but the above considerations do not, at least, contradict it. For, if the binding energy of the proton in the nucleus becomes comparable with  $m_U c^2$ , the range of interaction of the neutron with the former will increase considerable with the velocity of the neutron, so that the cross section will decrease slower in such case than in the case of hydrogen, i.e., free proton. Now the binding energy of the proton in  $C^{12}$ , which is estimated from the difference of masses of  $C^{12}$  and  $B^{11}$ , is

$$12,0036 - 11,0110 = 0,9926.$$

This corresponds to a binding energy 0,0152 in mass unit, being thirty times the electron mass. Thus in the case of carbon we can expect the effect observed by Bonner. The arguments are only tentative, other explanations being, of course, not excluded.

Next if  $\lambda < \frac{|\omega|}{c}$  or  $m_U c^2 < |W_N - W_P|$ ,  $\mu$  becomes pure imaginary and  $U$  expresses spherical undamped wave, implying that a quantum with energy greater than  $m_U c^2$  can be emitted in outer space by the transition of the heavy particle from neutron state to proton state, provided that  $|W_N - W_P| > m_U c^2$ .

The velocity of  $U$ -wave is greater but the group velocity is smaller than the light velocity  $c$ , as in the case of the electron wave.

The reason why such massive quanta, if they ever exist, are not yet discovered may be ascribed to the fact that the mass  $m_U$  is so large that condition  $|W_N - W_P| > m_U c^2$  is not fulfilled in ordinary nuclear transformation.

## § 4. Theory of $\beta$ – Disintegration

Hitherto we have considered only the interaction of  $U$ -quanta with heavy particles. Now, according to our theory, the quantum emitted when a heavy particle jumps from a neutron state to a proton state can be absorbed by a light particle which will then in consequence of energy absorption rise from a neutrino state of negative energy to an electron state of positive energy. thus an anti-neutrino and an electron are emitted simultaneously from the nucleus. Such intervention of a

massive quantum does not alter essentially the probability of  $\beta$ -disintegration, which has been calculated on the hypothesis of direct coupling of a heavy particle and a light particle, just as, in the theory of internal conversion of  $\gamma$ -ray, the intervention of the proton does not affect the final result. Our theory, therefore, does not differ essentially from Fermi's theory.

Fermi considered that an electron and a neutrino are emitted simultaneously from the radioactive nucleus, but this is formally equivalent to the assumption that a light particle jumps from a neutrino state of negative energy to an electron state of positive energy.

For, if the eigenfunctions of the electron and the neutrino be  $\Psi_k, \phi_k$  respectively, where  $k = 1, 2, 3, 4$ , a term of the form

$$-4\pi g' \sum_{k=1}^4 \tilde{\psi}_k \phi_k \quad (15)$$

should be added to the right hand side of the equation (5) for  $\tilde{U}$ , where  $g'$  is a new constant with the same dimension as  $g$ .

Now the eigenfunctions of the neutrino state with energy and momentum just opposite to those of the state  $\phi_k$  is given by  $\phi'_k = -\delta_{kl}\bar{\phi}_l$  and conversely  $\phi_k = \delta_{kl}\bar{\phi}_l$ , where

$$\delta = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix},$$

so that (15) becomes

$$-4\pi g' \sum_{k,l=1}^4 \tilde{\psi}_k \delta_{kl} \bar{\phi}'_l. \quad (16)$$

From equations (13) and (15), we obtain for the matrix element of the interaction energy of the heavy particle and the light particle an expression

$$gg' \int \dots \int \tilde{\nu}(\vec{r}_1) u(\vec{r}_1) \sum_{k=1}^4 \tilde{\psi}_k(\vec{r}_2) \phi_k(\vec{r}_2) \frac{e^{-\lambda r_{12}}}{r_{12}} d\nu_1 d\nu_2, \quad (17)$$

corresponding to the following double process: a heavy particle falls from the neutron state with the eigenfunction  $u(\vec{r})$  into the proton state with the eigenfunction  $\nu(\vec{r})$  and simultaneously a light particle jumps from the neutrino state  $\phi_k(\vec{r})$  of negative energy to the electron state  $\psi_k(\vec{r})$  of positive energy. In (17)  $\lambda$  is taken instead of  $\mu$ , since the difference of energies of the neutron state and the proton state, which is equal to the sum of the upper limit of the energy spectrum of  $\beta$ -rays and the proper energies of the electron and the neutrino, is always small compared with  $m_U c^2$ .

As  $\lambda$  is much larger than the wave numbers of the electron state and the neutrino state, the function  $\frac{e^{-\lambda r_{12}}}{r_{12}}$  can be regarded approximately as a  $\delta$ -function multiplied by  $\frac{4\pi}{\lambda^2}$  for the integrations with respect to  $x_2, y_2, z_2$ .

The factor  $\frac{4\pi}{\lambda^2}$  comes from

$$\int \int \int \frac{e^{-\lambda r_{12}}}{r_{12}} d\nu_2 = \frac{4\pi}{\lambda^2}.$$

Hence (17) becomes

$$\frac{4\pi gg'}{\lambda^2} \int \int \int \tilde{\nu}(\vec{r}) u(\vec{r}) \sum_k \tilde{\psi}_k(\vec{r}) \phi_k(\vec{r}) d\nu \quad (18)$$

or by (16)

$$\frac{4\pi gg'}{\lambda^2} \int \int \int \tilde{\nu}(\vec{r}) u(\vec{r}) \sum_{k,l} \tilde{\psi}(\vec{r}) \delta_{kl} \tilde{\phi}'_l(\vec{r}) d\nu, \quad (19)$$

which is the same as the expression (21) of Fermi, corresponding to the emission of a neutrino and an electron of positive energy states  $\phi'_k(\vec{r})$  and  $\psi_k(\vec{r})$ , except that the factor  $\frac{4\pi gg'}{\lambda^2}$  is substituted for Fermi's  $g$ .

Thus the result is the same as that of Fermi's theory, in this approximation, if we take

$$\frac{4\pi gg'}{\lambda^2} = 4 \times 10^{-50} \text{ cm}^3 \cdot \text{erg},$$

from which the constant  $g'$  can be determined. Taking, for example,  $\lambda = 5 \times 10^{12}$  and  $g = 2 \times 10^{-9}$ , we obtain  $g' \cong 4 \times 10^{-17}$ , which is about  $10^{-8}$  times as small as  $g$ .

This means that the interaction between the neutrino and the electron is much smaller than between the neutron and the proton so that the neutrino will be far more penetrating than the neutron and consequently more difficult to observe. The difference of  $g$  and  $g'$  may be due to the difference of masses of heavy and light particles.

## Summary

The interactions of elementary particles are described by considering a hypothetical quantum which has the elementary charge and the proper mass and which obeys Bose's statistics. The interaction of such a quantum with the heavy particle should be far greater than that with the light particle in order to account for the large interaction of the neutron and the proton as well as the small probability of  $\beta$ -disintegration.

Such quanta, if they ever exist and approach the matter close enough to be absorbed, will deliver their charge and energy to the latter. If, then, the quanta with negative charge come out in excess, The matter will be charged to a negative potential.

These arguments, of course, of merely speculative character, agree with the view that the high speed positive particles in the cosmic rays are generated by the electrostatic field of the earth, which is charged to a negative potential.

The massive quanta may also have some bearing on the shower produced by cosmic rays.



**LETTERS TO THE EDITOR**

*Prompt publication of brief reports of important discoveries in physics may be secured by addressing them to this department. Closing dates for this department are, for the first issue of the month, the eighteenth of the preceding month, for the second issue, the third of the month. Because of the late closing dates for the section no proof can be shown to authors. The Board of Editors does not hold itself responsible for the opinions expressed by the correspondents.*

**Communications should not in general exceed 600 words in length.**

P.A. Cerenkov  
The Physical Institute of the Academy of Sciences U.S.S.R., Moscow  
Received June 15, 1937

**Visible Radiation Produced by Electrons Moving in a Medium with Velocities Exceeding that of Light**

In a note published in 1934 [1] as well as in the subsequent publications [2] [3] [4] the present author reported his discovery of feeble visible radiation emitted by pure liquids under the action of fast electrons ( $\beta$ -particles of radioactive elements or Compton electrons liberated in liquids in the process of scattering of  $\gamma$ -rays). This radiation was a novel phenomenon, which could not be identified with any of the kinds of luminescence then known as the theory of luminescence failed to account for a number of unusual properties (insensitiveness to the action of quenching agents, anomalous polarization, marked spacial asymmetry, etc.) exhibited by the radiation in question. In 1934 the earliest results obtained in the experiments with  $\gamma$ -rays led S.I. Wawilow [5] to interpret the radiation observed as a result of the retardation of the Compton electrons liberated in liquids by  $\gamma$ -rays. A comprehensive quantitative theory subsequently advanced by I.M. Frank and I.E. Tamm

[6] afforded an exhaustive interpretation of all the peculiarities of the new phenomenon, including its most remarkable characteristic – the asymmetry.

According to their theory, an electron moving in a medium of refractive index  $n$  with a velocity exceeding that of light in the same medium ( $\beta > 1/n$ ) is liable to emit light which must be propagated in a direction forming an angle  $\theta$  with the path of the electron, this angle being determined by the equation:

$$\cos \theta = 1/\beta n, \quad (1)$$

where  $\beta$  is the ratio of the electron velocity to that of light in vacuum.

A successful experimental verification of formula (1) was only performed with water [4] for which, at the moment

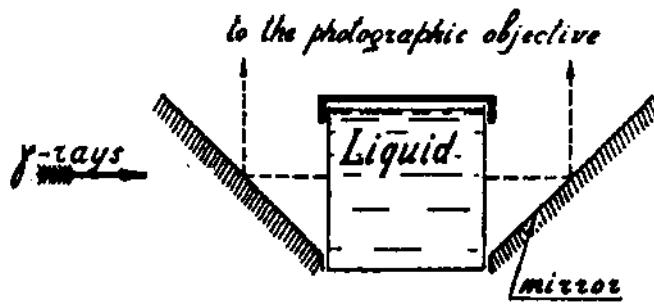


Figure 1: Arrangement of apparatus.

of publication of the above theory, data were already available which had been obtained by visual observations by the method of quenching [7] [8].

We recently performed additional experiments in which the intensity of radiation was recorded photographically, the records being taken simultaneously for all the angles  $\theta$  lying in a plane passing through the primary electron

beam. The liquid was placed in a cylindrical glass vessel with very thin walls, and the light emitted by the liquid was reflected by a conical mirror in an upward direction to the object glass of a photographic camera as indicated in Fig. 1. An approximately parallel beam of  $\gamma$ -rays, filtered through a 3-mm lead plate, fell on the liquid horizontally. The  $\gamma$ -radiation used was equivalent to that of 794 mg of radium. The considerable thickness of the lead screen, the large aperture of the object glass ( $f : 1.4$ ) and the long exposure (72 hours) ensured sufficient distinctness of the photographs.

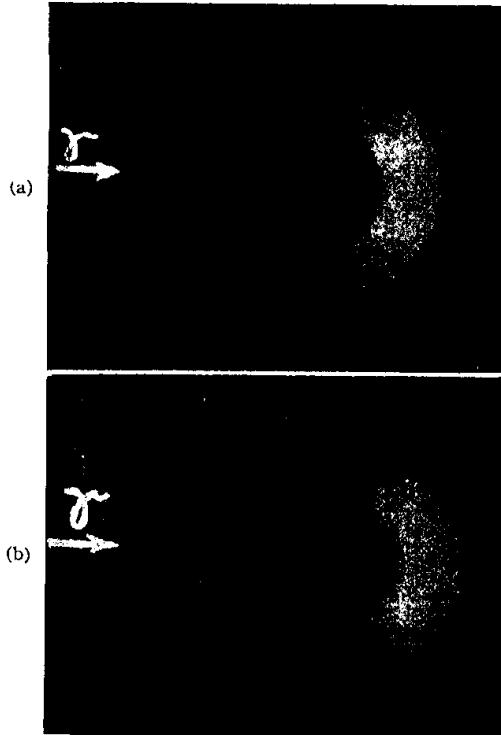


Figure 2: Photographs showing asymmetry of luminescence, (a) water,  $n = 1.337$ ; (b) benzene,  $n = 1.513$ .

The latter were obtained for ten different liquids. Two of the photographs taken (positive) are represented in Fig. 2. An examination of these photographs leads to the following conclusions:

- (1) In all the pure liquids investigated the radiation propagates mainly in the onward direction of the primary beam, the blackening of the negatives being only visible on part of the annular circle.
- (2) The area of the blackened sector increases with the refractive index of the liquids (see Fig. 2: (a)  $n = 1.337$  for water and (b)  $n = 1.513$  for benzene).
- (3) Each photograph exhibits two diffuse but clearly visible maxima of blackening, which are symmetrical with respect to the primary beam. Their

angular spacing increases with the refractive index of the liquids, and, to a first approximation, agrees with the values which might be expected according to Eq. (1). The absence of distinct maxima of blackening is undoubtedly associated with the difference in energy of the Compton electrons liberated from the molecules of the liquids by  $\gamma$ -rays, with the non-parallelism of these electrons and with the fact that the energy of each electron, moving in a liquid, gradually changes from the initial energy to zero.

All the results obtained are in good agreement with I.M. Frank and I.E. Tamm's theory of the coherent radiation of electrons moving in a medium [6].

P. A. CERENKOV

The Physical Institute of the Academy of Sciences of U.S.S.R., Moscow,  
June 15, 1937.

## References

- [1] Cerenkov, C.R. Ac. Sci. U.S.S.R. **8**, 451 (1934).
- [2] Cerenkov, C.R. Ac. Sci. U.S.S.R. **12** (3), 413 (1936).
- [3] Cerenkov. C.R. Ac. Sci. U.S.S.R. **14**, 102 (1937).
- [4] Cerenkov, C.R. Ac. Sci. U.S.S.R. **14**, 105 (1937).
- [5] Wawilow, C.R. Ac. Sci. U.S.S.R. **8**, 457 (1934).
- [6] Frank and Tamm, C.R. Ac. Sd. U.S.S.R. **14**, 109 (1937).
- [7] Bull. Ac. Sci. U.S.S.R. No. 7, 919 (1933).
- [8] E. Brumberg and S. Wawilow. C.R. Ac. Sci. U.S.S.R. **3**, 405 (1934)

**LETTERS TO THE EDITOR**

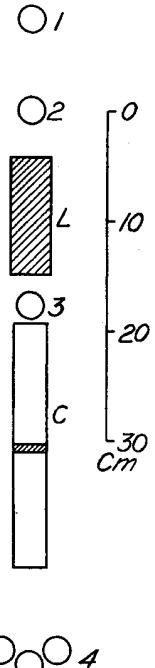
*Prompt publication of brief reports of important discoveries in physics may be secured by addressing them to this department. Closing dates for this department are, for the first issue of the month, the eighteenth of the preceding month, for the second issue, the third of the month. Because of the late closing dates for the section no proof can be shown to authors. The Board of Editors does not hold itself responsible for the opinions expressed by the correspondents.*

**Communications should not in general exceed 600 words in length.**

J.C. Street, E.C. Stevenson  
Research Laboratory Of Physics,  
Harvard University, Cambridge  
Received October 6, 1937

**New Evidence for the Existence of a Particle of Mass Intermediate Between the Proton and Electron**

Anderson and Neddermeyer [1] have shown that, for energies up to 300 and 400 Mev, the cosmic-ray shower particles have energy losses in lead plates corresponding to those predicted by theory for electrons. Recent studies of range [2] and energy loss [3] indicate that the singly occurring cosmic-ray corpuscles, even in the energy range below 400 Mev, are more penetrating than shower particles of corresponding magnetic deflection. Thus the natural assumptions have been expressed: the shower particles are electrons, the theory describing their energy losses is satisfactory, and the singly occurring particles are not electrons. The experiments cited above have shown from consideration of the specific ionization that the penetrating rays are not protons. The suggestion has been made that they are particles of electronic charge, and of mass intermediate between those of the proton and electron. If this is true, it should be possible to distinguish clearly such a particle from



O<sub>1</sub> O<sub>2</sub> O<sub>3</sub>

Figure 1: Geometrical arrangement of apparatus.

an electron or proton by observing its track density and magnetic deflection near the end of its range, although it is to be expected that the fraction of the total range in which the distinction can be made is very small. To examine this possibility experimentally we have used the arrangement of apparatus of Fig. 1. The three-counter telescope consisting of tubes 1, 2, and 3 and a lead filter  $L$  for removing shower particles, selects penetrating rays directed toward the cloud chamber  $C$  which is in a magnetic field of 3500 gauss. The type of track desired is one so near the end of its range as it enters the chamber that there is no chance of emergence below. In order to reduce the number of photographs of high energy particles, the tube group 4 was used as a cut-off counter with a circuit so arranged that the chamber would be set off only in those cases when a coincident discharge of counters 1, 2, and 3 was unaccompanied by a discharge of 4. The tripping of the cloud chamber valve was delayed about one sec. to facilitate determination of the drop count along a track. Because of geometrical imperfections

of the arrangement and of counter inefficiency the cut-off circuit prevented expansion for only  $\frac{3}{4}$  of the discharges of the telescope. At the present time



Figure 2: Track *A*.

1000 photos have been taken (equivalent to 4000 if the cut-off counter had not been used). Two tracks of interest, in that they have ionization densities definitely greater than usual, have been obtained: one *A* (see Fig. 2) is believed due to a proton and the other *B* (see Fig. 3) to a particle of mass approximately 130 times the rest mass of an electron.

Track *A* which terminated in the lead strip at the center of the chamber exhibited an ionization density 2.4 times as great as the usual thin tracks and an  $H\rho$  value approximately  $2 \times 10^6$  gauss cm in a direction to indicate a positive particle. Track *B* which passed out of the lighted region above the lead plate had an ionization density about six times as great as normal thin tracks (the ion density was too great to permit an accurate ion count) and an  $H\rho$  value of  $9.6 \times 10^4$  gauss cm. If it is assumed, as seems reasonable,

that the particle entered from above, the sign is negative. If it is taken that the ionization density varies inversely as the velocity squared, the rest mass of the particle in question is found to be approximately 130 times the rest mass of the electron. Because of uncertainty in the ion count this determination has a probable error of some 25 percent. In any case it does not seem possible to explain this track as due to a proton traveling up, for the observed  $H\rho$  value would indicate a proton of  $4.4 \times 10^5$  electron volts energy and therefore with a range of approximately one cm in the chamber. The track is clearly visible for 7 cm in the chamber.

The only possible objection to the conclusions reached above is that the bending of track *A* is largely due to distortion, but this is very unlikely, for the deflection is quite uniform and has a maximum value greater than ten times any distortions usually encountered in the thin tracks of high energy particles.

J.C. STREET  
E.C. STEVENSON

Research Laboratory of Physics,  
Harvard University,  
Cambridge, Massachusetts,  
October 6, 1937.

## References

- [1] Anderson and Neddermeyer, Phys. Rev. **50**, 263 (1936).
- [2] Street and Stevenson, Phys. Rev. **51**, 1005 (1937).
- [3] Neddermeyer and Anderson, Phys. Rev. **51**, 885 (1937).



Figure 3: Track *B*.



Figure 4: Photograph of the track of a penetrating particle of high energy for comparison with *A* and *B*.

## On Massive Neutron Cores

J. R. OPPENHEIMER AND G. M. VOLKOFF

*Department of Physics, University of California, Berkeley, California*

(Received January 3, 1939)

It has been suggested that, when the pressure within stellar matter becomes high enough, a new phase consisting of neutrons will be formed. In this paper we study the gravitational equilibrium of masses of neutrons, using the equation of state for a cold Fermi gas, and general relativity. For masses under  $\frac{1}{3}\odot$  only one equilibrium solution exists, which is approximately described by the nonrelativistic Fermi equation of state and Newtonian gravitational theory. For masses  $\frac{1}{3}\odot < m < \frac{3}{4}\odot$  two solutions exist, one stable and quasi-Newtonian, one more condensed, and unstable. For masses greater than  $\frac{3}{4}\odot$  there are no static equilibrium solutions. These results are qualitatively confirmed by comparison with suitably chosen special cases of the analytic solutions recently discovered by Tolman. A discussion of the probable effect of deviations from the Fermi equation of state suggests that actual stellar matter after the exhaustion of thermonuclear sources of energy will, if massive enough, contract indefinitely, although more and more slowly, never reaching true equilibrium.

### I. INTRODUCTION

FOR the application of the methods commonly used in attacking the problem of stellar structure<sup>1</sup> the distribution of energy sources and their dependence on the physical conditions within the star must be known. Since at the time of Eddington's original studies not much was known about the physical processes responsible for the generation of energy within a star, various mathematically convenient assumptions were made in regard to the energy sources, and these led to different star models (e.g. the Eddington model, the point source model, etc.). It was found that with a given equation of state for the stellar material many important properties of the solutions (such as the mass-luminosity law) were quite insensitive to the choice of assumptions about the distribution of energy sources, but were common to a wide range of models.

In 1932 Landau<sup>2</sup> proposed that instead of making arbitrary assumptions about energy sources chosen merely for mathematical convenience, one should attack the problem by first investigating the physical nature of the equilibrium of a given mass of material in which no energy is generated, and from which there is no radiation, presumably in the hope that such an

investigation would afford some insight into the more general situation where the generation of energy is taken into account. Although such a model gives a good description of a white dwarf star in which most of the material is supposed to be in a degenerate state with a zero point energy high compared to thermal energies of even  $10^7$  degrees, and such that the pressure is determined essentially by the density only and not by the temperature, still it would fail completely to describe a normal main sequence star, in which on the basis of the Eddington model the stellar material is nondegenerate, and the existence of energy sources and of the consequent temperature and pressure gradients plays an important part in determining the equilibrium conditions. The stability of a model in which the energy sources have to be taken into account is known to depend also on the temperature sensitivity of the energy sources and on the presence or absence of a time-lag in their response to temperature changes. However, if the view which seems plausible at present is adopted that the principal sources of stellar energy, at least in main sequence stars, are thermonuclear reactions, then the limiting case considered by Landau again becomes of interest in the discussion of what will eventually happen to a normal main sequence star after all the elements available for thermonuclear reactions are used up. Landau showed that for a model consisting of a cold degenerate Fermi gas there exist no stable equilibrium configurations for

<sup>1</sup> A. Eddington, *The Internal Constitution of the Stars* (Cambridge University Press, 1926); B. Strömgren, *Ergebn. Exakt. Naturwiss.* **16**, 465 (1937); Short summary in G. Gamow, *Phys. Rev.* **53**, 595 (1938).

<sup>2</sup> L. Landau, *Physik. Zeits. Sowjetunion* **1**, 285 (1932).

masses greater than a certain critical mass, all larger masses tending to collapse. For a mixture of electrons and nuclei in which on the average there are two protonic masses per electron Landau found the critical mass to be roughly  $1.5\odot$ , and in general the critical mass is inversely proportional to the square of the mass per particle obtained by spreading out the total mass over only those particles which essentially determine the pressure of the Fermi gas.

The possibility has been suggested<sup>3</sup> that in sufficiently massive stars after all the thermonuclear sources of energy, at least for the central material of the star, have been exhausted a condensed neutron core would be formed. The minimum mass for which such a core would be stable has been estimated by Oppenheimer and Serber,<sup>4</sup> who on taking into account some effects of nuclear forces give approximately  $0.1\odot$  as a reasonable minimum mass. The gradual growth of such a core with the accompanying liberation of gravitational energy is suggested by Landau as a possible source of stellar energy.

In this connection it seems of interest to ask whether this model of the final state of a star can be right for arbitrarily heavy stars, i.e., to investigate whether there is an upper limit to the possible size of such a neutron core. Landau's original result for a cold relativistically degenerate Fermi gas quoted above gives in the case of a neutron gas an upper limit of about  $6\odot$  beyond which the core would not be stable but would tend to collapse. Two objections might be raised against this result. One is that it was obtained on the basis of Newtonian gravitational theory while for such high masses and densities general relativistic effects must be considered. The other one is that the Fermi gas was assumed to be relativistically degenerate throughout the whole core, while it might be expected that on the one hand, because of the large mass of the neutron, the nonrelativistically degenerate equation of state might be more appropriate over the greater part of the core, and on the other hand the gravitational effect of the kinetic energy of the neutrons could not be neglected. The present

investigation seeks to establish what differences are introduced into the result if general relativistic gravitational theory is used instead of Newtonian, and if a more exact equation of state is used. A discussion of the general relativistic treatment of the equilibrium of spherically symmetric distributions of matter is first given, and then the special ideal case of a cold neutron gas is treated. A discussion of the results, and comparison with some results of Professor R. C. Tolman reported in an accompanying paper are given in the concluding sections.

## II. RELATIVISTIC TREATMENT OF EQUILIBRIUM

It is known<sup>5</sup> that the most general static line element exhibiting spherical symmetry may be expressed in the form

$$ds^2 = -e^\lambda dr^2 - r^2 d\theta^2 - r^2 \sin^2 \theta d\phi^2 + e^\nu dt^2, \quad (1)$$

$$\lambda = \lambda(r), \quad \nu = \nu(r).$$

If the matter supports no transverse stresses and has no mass motion, then its energy momentum tensor is given by<sup>6</sup>

$$T_1^1 = T_2^2 = T_3^3 = -p, \quad T_4^4 = \rho \quad (2)$$

where  $p$  and  $\rho$  are respectively the pressure and the macroscopic energy density measured in proper coordinates. With these expressions for the line element and for the energy momentum tensor, and with the cosmological constant  $\Lambda$  taken equal to zero, Einstein's field equations reduce to:<sup>7</sup>

$$8\pi p = e^{-\lambda} \left( \frac{\nu'}{r} + \frac{1}{r^2} \right) - \frac{1}{r^2}, \quad (3)$$

$$8\pi\rho = e^{-\lambda} \left( \frac{\lambda'}{r} - \frac{1}{r^2} \right) + \frac{1}{r^2}, \quad (4)$$

$$\frac{dp}{dr} \doteq -\frac{(p+\rho)}{2}\nu', \quad (5)$$

where primes denote differentiation with respect to  $r$ . These three equations together with the equation of state of the material  $\rho = \rho(p)$  de-

<sup>3</sup> G. Gamow, *Atomic Nuclei and Nuclear Transformations* (Oxford, 1936), second edition, p. 234. L. Landau, Nature **141**, 333 (1938) and others.

<sup>4</sup> J. R. Oppenheimer and R. Serber, Phys. Rev. **54**, 540 (1938).

<sup>5</sup> R. C. Tolman, *Relativity, Thermodynamics and Cosmology* (Oxford, 1934), pp. 239-241.

<sup>6</sup> R. C. Tolman, reference 5, p. 243.

<sup>7</sup> R. C. Tolman, reference 5, p. 244.

determine the mechanical equilibrium of the matter distribution as well as the dependence of the  $g_{\mu\nu}$ 's on  $r$ .

The boundary of the matter distribution is the value of  $r=r_b$  for which  $p=0$ , and such that for  $r < r_b$ ,  $p > 0$ . For  $r < r_b$  the solution depends on the equation of state of the material connecting  $p$  and  $\rho$ . For many equations of state a sharp boundary exists with a finite value of  $r_b$ .

In empty space surrounding the spherically symmetric distribution of matter  $p=\rho=0$ , and Schwarzschild's exterior solution is obtained:<sup>8</sup>

$$e^{-\lambda(r)} = 1 + A/r, \quad e^{\nu(r)} = B(1 + A/r). \quad (6)$$

The constants  $A$  and  $B$  are fixed by the requirement that at great distances away from the matter distribution the  $g_{\mu\nu}$ 's must go over into their weak-field form, i.e.,  $B=1$ ,  $A=-2m$  where  $m$  is the total Newtonian mass of the matter as calculated by a distant observer.<sup>9</sup>

Inside the boundary Eqs. (3), (4) and (5) may be rewritten as follows. Using the equation of state  $\rho=\rho(p)$  Eq. (5) may be immediately integrated.

$$\begin{aligned} \nu(r) &= \nu(r_b) - \int_0^{p(r)} \frac{2dp}{p+\rho(p)}, \\ e^{\nu(r)} &= e^{\nu(r_b)} \exp \left[ - \int_0^{p(r)} \frac{2dp}{p+\rho(p)} \right]. \end{aligned}$$

The constant  $e^{\nu(r_b)}$  is determined by making  $\nu'$  continuous across the boundary.

$$e^{\nu(r)} = \left( 1 - \frac{2m}{r_b} \right) \exp \left[ - \int_0^{p(r)} \frac{2dp}{p+\rho(p)} \right]. \quad (7)$$

Thus  $e^{\nu}$  is known as a function of  $r$  if  $p$  is known as a function of  $r$ . Further in Eq. (4) introduce a new variable

$$u(r) = \frac{1}{2}r(1 - e^{-\lambda}) \quad \text{or} \quad e^{-\lambda} = 1 - 2u/r. \quad (8)$$

Then Eq. (4) becomes:

$$du/dr = 4\pi\rho(p)r^2. \quad (9)$$

In Eq. (3) replace  $e^{-\lambda}$  by its value from (8) and  $\nu'$  by its value from (5). It becomes:

$$\frac{dp}{dr} = -\frac{p+\rho(p)}{r(r-2u)} [4\pi pr^3 + u]. \quad (10)$$

<sup>8</sup> R. C. Tolman, reference 5, p. 203.

<sup>9</sup> R. C. Tolman, reference 5, pp. 203 and 207.

Equations (9) and (10) form a system of two first-order equations in  $u$  and  $p$ . Starting with some initial values  $u=u_0$ ,  $p=p_0$  at  $r=0$ , the two equations are integrated simultaneously to the value  $r=r_b$  where  $p=0$ , i.e., until the boundary of the matter distribution is reached. The value of  $u=u_b$  at  $r=r_b$  determines the value of  $e^{\nu(r_b)}$  at the boundary, and this is joined continuously across the boundary to the exterior solution, making

$$u_b = \frac{r_b}{2} [1 - e^{-\lambda(r_b)}] = \frac{r_b}{2} \left[ 1 - \left( 1 - \frac{2m}{r_b} \right) \right] = m.$$

Thus the mass of this spherical distribution of matter as measured by a distant observer is given by the value  $u_b$  of  $u$  at  $r=r_b$ .

The following restrictions must be made on the choice of  $p_0$  and  $u_0$ , the initial values of  $p$  and  $u$  at  $r=0$ :

(a) In accordance with its physical meaning as pressure,  $p_0 \geq 0$ .

(b) From Eq. (8) it is seen that for all finite values of  $e^{-\lambda}$ ,  $u_0=0$ . Since  $g_{11}=-e^\lambda$  must never be positive,  $u_0 \leq 0$  for infinite values of  $e^{-\lambda}$  at the origin. However, it may be shown that of all the finite values of  $p_0$  at the origin  $p_0=0$  is the only one compatible with a negative value of  $u_0$ , and that for equations of state of the type occurring in this problem even this possibility is excluded, so that  $u_0$  must vanish.<sup>10</sup>

(c) A special investigation for any particular equation of state must be made to see whether solutions exist in which  $0 \leq u_0 \leq -\infty$  and  $p \rightarrow \infty$  as  $r \rightarrow 0$ .

### III. PARTICULAR EQUATIONS OF STATE

The above arguments show that Eqs. (9) and (10) together with a given equation of state completely determine the distribution of matter.

<sup>10</sup> This can be seen from the following argument. Having chosen some particular value of  $p_0$  one may usually represent the equation of state in that pressure range by  $\rho=Kp^s$  with some appropriate value of  $s$ . Using this equation of state and taking the approximate form of Eq. (10) near the origin for the case  $u_0 < 0$ , and finite  $p_0$ , one obtains:

$$\frac{dp}{dr} = \frac{p+\rho(p)}{2r} = \frac{p+Kp^s}{2r}.$$

Integration of this equation shows that for  $s < 1$   $p_0 \geq 0$  can not be satisfied, and for  $s \geq 1$  only the value  $p_0=0$  is possible. For the equations of state used in this problem always  $s < 1$  holds. It may also be noted that the above equation together with Eq. (7) show that  $e^{\nu(r)} \rightarrow \infty$  as  $r \rightarrow 0$ .

The assumption  $\rho = \text{const.}$ ,  $u_0 = 0$  makes it possible to integrate Eqs. (9) and (10) explicitly and leads to Schwarzschild's interior solution.<sup>11</sup> Other matter distributions corresponding to other equations of state are given by Professor Tolman in an accompanying paper.

If the matter is taken to consist of particles of rest mass  $\mu_0$  obeying Fermi statistics, and their thermal energy<sup>12</sup> and all forces between them are neglected, then it may be shown that a parametric form for the equation of state is:<sup>13</sup>

$$\rho = K(\sinh t - t), \quad (11)$$

$$p = \frac{1}{3}K(\sinh t - 8 \sinh \frac{1}{2}t + 3t), \quad (12)$$

$$\text{where } K = \pi\mu_0^4 c^5 / 4h^3 \quad (13)$$

$$\text{and } t = 4 \log \left( \frac{\hat{p}}{\mu_0 c} + \left[ 1 + \left( \frac{\hat{p}}{\mu_0 c} \right)^2 \right]^{\frac{1}{2}} \right), \quad (14)$$

where  $\hat{p}$  is the maximum momentum in the Fermi distribution and is related to the proper particle density  $N/V$  by

$$\frac{N}{V} = \frac{8\pi}{3h^3} \hat{p}^3. \quad (15)$$

Substituting the above expressions for  $p$  and  $\rho$  into Eqs. (9) and (10) one gets:

$$\frac{du}{dr} = 4\pi r^2 K(\sinh t - t), \quad (16)$$

$$\frac{dt}{dr} = -\frac{4}{r(r-2u)} \frac{\sinh t - 2 \sinh \frac{1}{2}t}{\cosh t - 4 \cosh \frac{1}{2}t + 3} \\ \times [(4/3)\pi Kr^3(\sinh t - 8 \sinh \frac{1}{2}t + 3t) + u]. \quad (17)$$

<sup>11</sup> R. C. Tolman, reference 5, pp. 246-247.

<sup>12</sup> The condition for thermal equilibrium in a static gravitational field is given by Tolman (reference 5, p. 318) as  $T_0(g_{44})^{\frac{1}{2}} = \text{const.}$  where  $T_0$  is the proper temperature. The equilibrium state of a matter distribution which no longer radiates appreciably corresponds to a low surface temperature  $T_0$ . If  $g_{44}$  is everywhere finite, then  $T_0$  will be small throughout the matter distribution. For those singular solutions in which  $g_{44}$  vanishes at the origin it is conceivable that the central temperature may be high. However, on the one hand from Eq. (7) it is seen that the vanishing of  $g_{44}$  at the origin corresponds to infinite central pressure, and in this limit the equation of state given below reduces to  $\rho = 3p$  so that temperature introduces no radically new effects, and on the other hand zero values of  $g_{44}$  indicate the slowing down of all physical processes near the origin and thus may correspond to nonstatic solutions describing states which have not yet attained equilibrium, and which are not discussed in this paper.

<sup>13</sup> Cf. S. Chandrasekhar, Monthly Notices of R.A.S. 95, 222 (1935), but introduce *energy* density in place of his *mass* density.

These equations are to be integrated from the values  $u=0$ ,  $t=t_0$  at  $r=0$  to  $r=r_b$  where  $t_b=0$  (which makes  $p=0$ ), and  $u=u_b$ .

A note must be made of the units employed in these equations. Eqs. (3), (4) and (5) from which (16) and (17) are derived are stated in relativistic units,<sup>14</sup> i.e., such that  $c=1$ ,  $G=1$  ( $c$  is the velocity of light,  $G$  is the gravitational constant). This determines the unit of time and the unit of mass in terms of a still arbitrary unit of length. The unit of length is now fixed by the requirement that  $K=1/4\pi$ . Eqs. (16) and (17) now become:

$$\frac{du}{dr} = r^2(\sinh t - t), \quad (18)$$

$$\frac{dt}{dr} = -\frac{4}{r(r-2u)} \frac{\sinh t - 2 \sinh \frac{1}{2}t}{\cosh t - 4 \cosh \frac{1}{2}t + 3} \\ \times [\frac{1}{3}r^3(\sinh t + 8 \sinh \frac{1}{2}t + 3t) + u]. \quad (19)$$

The unit of length has been fixed to be

$$a = \frac{1}{\pi} \left( \frac{h}{\mu_0 c} \right)^{\frac{1}{3}} \frac{c}{(\mu_0 G)^{\frac{1}{3}}},$$

while the unit of mass is

$$b = \frac{c^2}{G} a = \frac{1}{\pi} \left( \frac{h}{\mu_0 c} \right)^{\frac{1}{3}} \frac{c^3}{(\mu_0 G^3)^{\frac{1}{3}}}.$$

For a neutron gas  $a = 1.36 \times 10^6$  cm,  $b = 1.83 \times 10^{34}$  g. The general character of the solution is seen to be independent of the mass of the neutron which determines only the scale of the result.

No way was found to carry out the integration analytically, so Eqs. (18) and (19) were integrated numerically for several finite values of  $t_0$ . For all these cases  $u_0$  was taken to be equal to zero, since the equation of state near the origin for finite  $t_0$  behaves like  $\rho(p) = Kp^s$ ,  $s < 1$ . The first four entries in Table I were thus obtained.

For  $t_0 \rightarrow \infty$  Eqs. (18) and (19) may be replaced by their asymptotic expressions:

$$du/dr = \frac{1}{2}r^2 e^t, \quad (20)$$

$$\frac{dt}{dr} = -\frac{4}{r(r-2u)} \left[ \frac{r^3}{6} e^t + u \right]. \quad (21)$$

<sup>14</sup> R. C. Tolman, reference 5, pp. 201-202.

An exact solution<sup>15</sup> of these equations is:

$$e^t = 3/7r^2, \quad u = 3r/14, \quad (22)$$

which corresponds to  $t_0 = \infty$ ,  $u_0 = 0$ . A careful examination of Eqs. (20) and (21) shows that there are no other solutions corresponding to  $t_0 = \infty$ ,  $0 \leq u_0 \leq -\infty$ . The exact solution (22) of the approximate equations (20) and (21) was taken out to that value of  $r$  where  $t=6$  (the approximation in the form of Eq. (20) and (21) is quite good for  $t \geq 6$ ), and then the integration of the exact equations (18) and (19) was carried out numerically to  $r=r_b$  where  $t=0$ . This gave the last entry in Table I.

It is of interest to ask whether perhaps a finite gravitational mass might correspond to an infinite number of particles, and an infinite gravitational binding energy. It may be seen that this is not the case by the following argument. Although the proper particle density becomes infinite when the central pressure becomes infinite, still it remains integrable, so that the total number of particles always remains finite. The element of proper volume of a spherical shell is  $4\pi e^{\lambda/2} r^2 dr$ . As the solution of the approximate equations shows in the neighborhood of the origin:

$$e^{\lambda/2} = \left(1 - \frac{2u}{r}\right)^{-\frac{1}{2}} = \left(1 - \frac{3}{7}\right)^{-\frac{1}{2}} = \left(\frac{7}{4}\right)^{\frac{1}{2}},$$

$N/V \propto \hat{p}^3 \propto e^{3t/4} \propto 1/r^{\frac{3}{2}}$  for large  $t$  and  $\hat{p}$ ;

$$\therefore N \propto \int_0^r \frac{r^2}{r^{\frac{3}{2}}} dr \propto r^{\frac{1}{2}} \text{ near the origin.}$$

*A fortiori* the number of particles is finite for nonsingular solutions.

<sup>15</sup> This solution is a limiting form of the solutions V, VI given by Tolman in the accompanying paper.

For very small values of  $t$  the equation of state (11), (12) reduces to  $\hat{p}=K\rho^{5/3}$  and  $\hat{p} \propto t$ . Using this equation of state and Newtonian gravitational theory (which is expected to give a good result for small masses and densities), one finds that  $\hat{p} \propto m^{\frac{5}{3}}$ , or that  $m \propto t^{\frac{3}{5}}$ . Fig. 1 gives a schematic plot of the dependence of  $m$  on  $t_0$  for the case that the elementary particles are neutrons. The mass  $m$  is plotted in units of sun's mass ( $2 \times 10^{33}$  g) against  $\tan^{-1} t_0$ . The curve near the origin is dotted since, as has been already pointed out, a neutron core with a mass less than about  $0.1\odot$  will disintegrate into nuclei and electrons.

The striking feature of the curve is that the mass increases with increasing  $t_0$  until a maximum is reached at about  $t_0=3$ , after which the curve drops until a value roughly  $\frac{1}{3}\odot$  is reached for  $t_0=\infty$ . In other words no static solutions at all exist for  $m > \frac{3}{4}\odot$ , two solutions exist for all  $m$  in  $\frac{3}{4}\odot > m > \frac{1}{3}\odot$ , and one solution exists for all  $m < \frac{1}{3}\odot$ .

Some insight into this situation may be gained from the following considerations. In the non-relativistic polytrope solutions of Emden<sup>16</sup> the equation of state was assumed to be  $\hat{p}=K\rho^\gamma = K\rho^{1+1/n}$ . Solutions which at first sight seem to be quite satisfactory (i.e., giving a finite mass within a finite radius) were found for values of  $n < 5$  or  $\gamma > 6/5$ . But Landau<sup>2</sup> pointed out that although these solutions in every case give an equilibrium configuration, they do not in every case give stable equilibrium. Thus, unless  $\gamma \geq 4/3$  the equilibrium configuration is unstable. This may be seen from the following rough calculation. The gravitational part of the free energy of the system is negative and proportional to  $\rho^{\frac{5}{3}}$  where

<sup>16</sup> Emden, *Gaskugeln* (1907), or cf. *Handbuch der Astrophys.* Vol. 3, p. 186.

TABLE I. Mass, radius and neutron density for various values of  $t_0$ .

$t_0$	MASS		RADIUS		$(\frac{\hat{p}}{\mu_0 c})_{r=0}$	$(\frac{N}{V})_{r=0}$ NEUTRONS CM <sup>3</sup>
	IN UNITS OF Eqs. (18), (19)	IN UNITS OF $\odot$ FOR NEUTRONS	IN UNITS OF Eqs. (18), (19)	IN KILOMETERS, FOR NEUTRONS		
1	0.033	0.30	1.55	21.1	0.25	$0.062 \times 10^{39}$
2	0.066	0.60	0.98	13.3	0.52	$0.56 \times 10^{39}$
3	0.078	0.71	0.70	9.5	0.82	$2.2 \times 10^{39}$
4	0.070	0.64	0.50	6.8	1.17	$6.4 \times 10^{39}$
$\infty$	0.037	0.34	0.23	3.1	$\infty$	$\infty$

$\rho$  is an appropriate average density (Newtonian gravitational theory is used). The part of the free energy caused by compression is proportional to  $\int p dv$ , and hence to  $\rho^{\gamma-1}$  ( $\gamma \neq 1$ ). Thus

$$F = -a\rho^{\frac{1}{\gamma}} + b\rho^{\gamma-1}.$$

Polytrope solutions exist for both  $\gamma = 5/3 (> 4/3)$ , i.e., for  $n = 3/2$  and for  $\gamma = 5/4 (< 4/3$ , but  $> 6/5)$ , i.e., for  $n = 4$ , but as may be seen from the schematic plot of the free energy curves in Fig. 2, the former corresponds to stable equilibrium and the latter to unstable equilibrium.

In the present relativistic calculations the results for small masses and small central densities and pressures (small values of  $t_0$ ), as was already mentioned above, may be expected to agree quite closely with nonrelativistic calculations with the equation of state  $p = K\rho^{5/3}$ . Since  $\rho$  is a monotonic function of  $t$ , the curves of free energy against  $t_0$  for fixed total number of particles, and thus for a fixed  $M_0$  (gravitational mass at zero density; the gravitational mass will vary somewhat along a curve of constant particle number, as the density increases), will for small masses have the same general character as the curves of free energy against some average density in the nonrelativistic case (cf. the curve for  $\gamma = 5/3$  in Fig. 2). Then as the number of particles is increased the character of the free energy curves must change in order to admit the possibility of a second equilibrium position. Since the free energy must be a continuous function of  $t_0$ , and since we know from nonrelativistic calculations that for small masses (and low densities) we have a position of stable equilibrium (a minimum in the free energy curve) we can conclude that the second equilibrium position corresponds either to a maximum or to an inflection point in the free energy curve (and certainly not to a minimum). Fig. 3 gives a schematic plot of free energy against  $t_0$  for different values of  $M_0$  which would explain the existence of one equilibrium position for small masses, two for intermediate masses, and none for large masses. The masses marked on the curves are the actual gravitational masses corresponding to the equilibrium points of the critical free energy curves separating the solutions into the three types mentioned above.

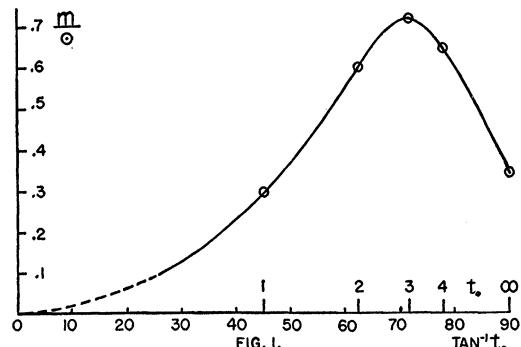


FIG. 1. Dependence of  $m$  on  $t_0$  for neutrons.

#### IV. DISCUSSION—RELATION TO TOLMAN'S SOLUTIONS

Before we study the physical implications of these results, we may try to show how their qualitative features may be obtained from the analytic solutions recently discovered by Professor Tolman.<sup>17</sup> This will also help us to understand the probable effect of alterations in the equation of state of the neutron gas at high densities.

On the one hand Tolman's solution IV, discussed in §6 of his paper, enables us to understand the existence of a limiting mass for static solutions and to give an estimate of its magnitude; on the other hand Tolman's solution VI, discussed in §8 (and less directly solution V), has for  $n = \frac{1}{2}$ , very much the character of our singular solution for  $t_0 \rightarrow \infty$ , and, with appropriately chosen constants, gives a mass of the same order of magnitude as we have found.

Tolman's solution IV is nonsingular, and corresponds to the quadratic equation of state (6.5) of his paper:

$$\frac{(\rho_c - \rho)^2}{8\rho_c + \rho_c} - 5(\rho_c - \rho) + \rho_c - \rho = 0 \quad (\text{Tolman, 6.5}),$$

where  $\rho_c$  and  $\rho_c$  are the central density and pressure. From Eqs. (6.4), (6.6) and (6.9) of Tolman's paper the mass corresponding to this solution is given in terms of  $\rho_c$  and  $\rho_c$  by

$$m = 4 \left( \frac{\rho_c}{\rho_c + 3\rho_c} \right)^{\frac{1}{2}} \left[ \frac{8\pi}{3} (\rho_c - 3\rho_c) \right]^{-\frac{1}{2}}. \quad (23)$$

<sup>17</sup> We are very much indebted to Professor Tolman for letting us see these results before publication, and for helpful discussions of them.

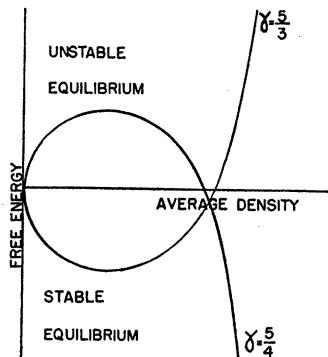


FIG. 2. Free energy as a function of average density.

If  $\rho_c$  and  $p_c$  are now themselves connected by the Fermi equation of state (11), (12), then  $p_c \propto \rho_c^{5/3}$  as  $\rho_c \rightarrow 0$ , and  $\rho_c - 3p_c \sim 0(\rho_c^{1/3})$  as  $\rho_c \rightarrow \infty$ , and  $m$  is seen to have a maximum value. For values of  $\rho_c$  corresponding to this maximum the equation of state (Tolman 6.5) does not differ qualitatively from the Fermi equation of state (11), (12), as may be seen by comparing for the two solutions the values of  $d \ln p/d \ln \rho$ ; and the maximum mass in fact turns out from (23) to be  $\sim 0.4\odot$ , agreeing in order of magnitude with our value of  $\sim 0.7\odot$ .

Tolman's solution V, with  $n = \frac{1}{2}$ ,  $R \rightarrow \infty$ , and his solution VI, with  $n = \frac{1}{2}$ ,  $B/A \rightarrow 0$ , are just our solution (22) corresponding to the equation of state  $\dot{p} = \frac{1}{3}\rho$ , a unique unstable singular solution. For solution V, with  $n = \frac{1}{2}$  and finite  $R$ , the pressure differs from  $\frac{1}{3}\rho$  by terms of the order of  $\rho^{-1/6}$ ; however, for VI with  $n = \frac{1}{2}$  and finite  $B/A$ , for large  $\rho$ ,  $\rho - 3p = \text{const. } \rho^{\frac{1}{3}}$ , which is just the behavior of a highly compressed Fermi gas. Using for the mass of this solution

$$m = A/42B \quad (\text{Tolman, 8.9})$$

and adjusting the ratio  $B/A$  to make the equation of state of VI, i.e.,

$$p = \frac{1}{3}\rho \frac{1 - 9(B/A)(3/56\pi)^{\frac{1}{3}}\rho^{-\frac{1}{3}}}{1 - (B/A)(3/56\pi)^{\frac{1}{3}}\rho^{-\frac{1}{3}}} \quad (\text{Tolman, 8.5})$$

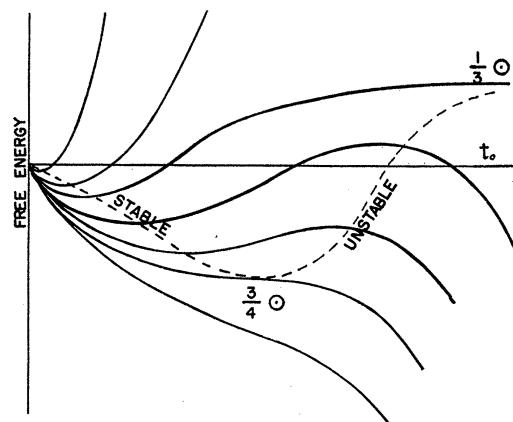
agree to terms of order  $\rho^{\frac{1}{3}}$  with (11), (12), we get  $B/A = (7/3)^{\frac{1}{3}}$ , and  $m \sim (1/7)\odot$ , to compare with the value of  $\frac{1}{3}\odot$  which the Fermi equation gives.

These necessarily somewhat rough comparisons may thus serve to give an idea of the analytic

character of the solutions corresponding to maximum mass and to maximum (infinite) central density which we obtained above.

#### V. DISCUSSION—APPLICATION TO STELLAR MATTER

We have seen that for a cold neutron core there are no static solutions, and thus no equilibrium, for core masses greater than  $m \sim 0.7\odot$ . The corresponding maximum mass  $M_0$  before collapse is some ten percent greater than this. Since neutron cores can hardly be stable (with respect to formation of electrons and nuclei) for masses less than  $\sim 0.1\odot$ , and since, even after thermonuclear sources of energy are exhausted, they will not tend to form by collapse of ordinary matter for masses under  $1.5\odot$  (Landau's limit), it seems unlikely that static neutron cores can play any great part in stellar evolution,<sup>18</sup> and the question of what happens, after energy sources are exhausted, to stars of mass greater than  $1.5\odot$  still remains unanswered. It should be observed that for the critical solution with  $m \sim 0.7\odot$  the potentials  $g_{\mu\nu}$  are nowhere singular, and that in particular such a core does not tend to "protect itself" from the addition of further matter by the vanishing of  $g_{44}$  at the boundary. There would then seem to be only two answers possible to the question of the "final" behavior of very massive stars: either the equation of

FIG. 3. Schematic plot of free energy as a function of  $t_0$ .

<sup>18</sup> The mass of the shell of ordinary (but dense) matter surrounding the core must be small for cores much more massive than the lightest core stable with respect to disintegration into electrons and nuclei.

state we have used so far fails to describe the behavior of highly condensed matter that the conclusions reached above are qualitatively misleading, or the star will continue to contract indefinitely, never reaching equilibrium. Both alternatives require serious consideration.

The central density in the "critical" core is even higher than nuclear density, so that our extrapolation of the Fermi equation of state can hardly rest on a very sure basis. Under these conditions the disintegration of neutrons, either into protons and electrons, or into mesotrons, will be energetically unfavorable and will not occur. And the relatively weak attractive forces which are known to act between neutrons will facilitate, and not prevent, the collapse of the core. If, however, under extreme compression, phenomena occurred which have the effect of repulsive forces, i.e., of raising the pressure for a given density above the value given by the Fermi equation of state, this could tend to prevent the collapse.

Such repulsive forces, even if they exist, will hardly make possible static solutions for arbitrarily large amounts of matter. For at low densities they cannot appreciably affect the equation of state, so that the dimensions of the core will necessarily be finite, and so will be the gravitational mass  $m$  of the core

$$m = \frac{1}{2}r_b(1 - e^{-\lambda b}) \quad (\text{Tolman, 5.5}).$$

Nor can the mass  $M_0$  before collapse be infinite. For this to be true we should have to have a singular solution. But the effect of repulsive forces can for high density at most be to make  $3p$  even more nearly equal to  $\rho$  than for the Fermi equation of state; and for  $\rho = 3p$ , as has been remarked above, and as is also suggested by Tolman's solutions V and VI, the *only* singular static solution is (22), for which the total particle number is finite.

We may obtain an extreme limit on the increase in the limiting mass which strong repulsive forces at high densities could give, by the following simple argument. For  $\rho < 10^{15}$  g/cm<sup>3</sup> these forces can hardly be important. Let us assume that for  $\rho \geq 10^{15}$ , they have the extreme

effect of making  $p = \frac{1}{3}\rho$ . Then the mass of a sphere for which this equation of state holds down to  $\rho = 10^{15}$ , and for which  $p$  falls rapidly as  $\rho \rightarrow 0$ , is given by our solution (22), and is of the order of  $\odot$ . It seems likely that our limit of  $\sim 0.7\odot$  is near the truth.

This argument is based on the requirement that even for arbitrarily high densities,  $\rho - 3p$  shall not be negative; and this is in turn closely related to the positive definite character of the (proper) energy density of neutrons and of the fields of force (apart from gravitation) associated with them. It seems probable that if  $p$  could be very much greater than  $\frac{1}{3}\rho$ , static solutions of arbitrarily large mass could be found.<sup>19</sup>

From this discussion it appears probable that for an understanding of the long time behavior of actual heavy stars a consideration of non-static solutions must be essential. Among all (spherical) nonstatic solutions one would hope to find some for which the rate of contraction, and in general the time variation, become slower and slower, so that these solutions might be regarded, not as equilibrium solutions, but as quasi-static. Some reason for this we may see in the following argument: for large enough mass the core will collapse; near the center the density and pressure will grow, and  $g_{44} = e^\nu$  will be small (cf. Eq. (7)); and as  $e^\nu$  grows smaller, all processes will, as seen by an outside observer, slow down in the central region. Formally one sees this, in the occurrence, in Einstein's equations, of products of the form

$$e^{-\nu} \frac{d^2\lambda}{dt^2}, \quad e^{-\nu} \left( \frac{d\lambda}{dt} \right)^2, \quad e^{-\nu} \frac{d\lambda}{dt} \frac{d\nu}{dt}.$$

For high enough central densities it is no longer justified to neglect even a very slow time variation; and the singular solutions which presumably represent very massive neutron cores cannot be obtained unless this is taken into account. These solutions are now being investigated.

---

<sup>19</sup> Thus for  $\rho = \text{const.}$  there is a class of singular static solutions, for which  $p \sim k/r^2$ , and which would seem to lead, for  $K \rightarrow \infty$ , to infinite masses, and which one of us (G.M.V.) hopes to discuss in detail elsewhere.

## Forces in Molecules

R. P. FEYNMAN

*Massachusetts Institute of Technology, Cambridge, Massachusetts*

(Received June 22, 1939)

Formulas have been developed to calculate the forces in a molecular system directly, rather than indirectly through the agency of energy. This permits an independent calculation of the slope of the curves of energy *vs.* position of the nuclei, and may thus increase the accuracy, or decrease the labor involved in the calculation of these curves. The force on a nucleus in an atomic system is shown to be just the classical electrostatic force that would be exerted on this nucleus by other nuclei and by the electrons' charge distribution. Qualitative implications of this are discussed.

MANY of the problems of molecular structure are concerned essentially with forces. The stiffness of valence bonds, the distortions in geometry due to the various repulsions and attractions between atoms, the tendency of valence bonds to occur at certain definite angles with each other, are some examples of the kind of problem in which the idea of force is paramount.

Usually these problems have been considered through the agency of energy, and its changes with changing configuration of the molecule. The reason for this indirect attack through energy, rather than the more qualitatively illuminating one, by considerations of force, is perhaps twofold. First it is probably thought that force is a quantity that is not easily described or calculated by wave mechanics, while energy is, and second, the first molecular problem to be solved is the analysis of band spectra, strictly a problem of energy as such. It is the purpose of this paper to show that forces are almost as easy to calculate as energies are, and that the equations are quite as easy to interpret. In fact, all forces on atomic nuclei in a molecule can be considered as purely classical attractions involving Coulomb's law. The electron cloud distribution is prevented from collapsing by obeying Schrödinger's equation. In these considerations the nuclei are considered as mass points held fixed in position.

A usual method of calculating interatomic forces runs somewhat as follows.

For a given, fixed configuration of the nuclei, the energy of the entire system (electrons and nuclei) is calculated. This is done by the variation method or other perturbation schemes. This

entire process is repeated for a new nuclear position, and the new value of energy calculated. Proceeding in this way, a plot of energy *vs.* position is obtained. The force on a nucleus is of course the slope of this curve.

The following method is one designed to obtain the forces at a given configuration, when only the configuration is known. It does not require the calculations at neighboring configurations. That is, it permits a calculation of the slope of the energy curve as well as its value, for any particular configuration. It is to be emphasized that this allows a considerable saving of labor of calculations. To obtain force under the usual scheme the energy needs to be calculated for two or more different and neighboring configurations. Each point requires the calculation of the wave functions for the entire system. In this new method, only one configuration, the one in question, need have its wave functions computed in detail. Thus the labor is considerably reduced. Because it permits one to get an independent value of the slope of the energy curve, the method might increase the accuracy in the calculation of these curves, being especially helpful in locating the normal separation, or position of zero force.

In the following it is to be understood that the nuclei of the atoms in the molecule, or other atomic system, are to be held fixed in position, as point charges, and the force required to be applied to the nuclei to hold them is to be calculated. This will lead to two possible definitions of force in the nonsteady state, for then the energy is not a definite quantity, and the slope of the energy curve shares this indefinite-

ness. It will be shown that these two possible definitions are exactly equivalent in the steady-state case, and, of course, no ambiguity should arise there.

Let  $\lambda$  be one of any number of parameters which specify nuclear positions. For example,  $\lambda$  might be the  $x$  component of the position of one of the nuclei. A force  $f_\lambda$  is to be associated with  $\lambda$  in such a way that  $f_\lambda d\lambda$  measures the virtual work done in displacing the nuclei through  $d\lambda$ . This will define the force only when the molecule is in a steady state, of energy  $U$ , for then we can say  $f_\lambda = -\partial U / \partial \lambda$ . In the non-steady-state case we have no sure guide to a definition of force. For example, if  $\bar{U} = \int \psi^* H \psi dv$  be the average energy of the system of wave function  $\psi$  and Hamiltonian  $H$ , we might define

$$f'_\lambda = -\partial(\bar{U})/\partial\lambda. \quad (1)$$

Or again, we might take  $f_\lambda$  to be the average of  $-\partial H / \partial \lambda$  or

$$f_\lambda = -\left\langle \frac{\partial H}{\partial \lambda} \right\rangle_{av} = -\int \psi^* \frac{\partial H}{\partial \lambda} \psi dv. \quad (2)$$

We shall prove that under steady-state conditions, both these definitions of force become exactly equivalent, and equal to  $-\partial U / \partial \lambda$ , the slope of the energy curve. Since (2) is simpler than (1) we can define force by (2) in general. In particular, it gives a simple expression for the slope of the energy curve.

Thus we shall prove, when  $H\psi = U\psi$  and  $\int \psi \psi^* dv = 1$  that,

$$\frac{\partial U}{\partial \lambda} = \int \psi^* \frac{\partial H}{\partial \lambda} \psi dv.$$

Now

$$U = \int \psi^* H \psi dv,$$

whence,

$$\frac{\partial U}{\partial \lambda} = \int \psi^* \frac{\partial H}{\partial \lambda} \psi dv + \int \frac{\partial \psi^*}{\partial \lambda} H \psi dv + \int \psi^* H \frac{\partial \psi}{\partial \lambda} dv.$$

Since  $H$  is a self-adjoint operator,

$$\int \psi^* H \frac{\partial \psi}{\partial \lambda} dv = \int \frac{\partial \psi}{\partial \lambda} H \psi^* dv.$$

But  $H\psi = U\psi$  and  $H\psi^* = U\psi^*$  so that we can write,

$$\frac{\partial U}{\partial \lambda} = \int \psi^* \frac{\partial H}{\partial \lambda} \psi dv + U \int \frac{\partial \psi^*}{\partial \lambda} \psi dv + U \int \frac{\partial \psi}{\partial \lambda} \psi^* dv.$$

These last two terms cancel each other since their sum is,

$$U \frac{\partial}{\partial \lambda} \int \psi^* \psi dv = U \frac{\partial}{\partial \lambda} (1) = 0.$$

Whence

$$\frac{\partial U}{\partial \lambda} = \int \psi^* \frac{\partial H}{\partial \lambda} \psi dv$$

in the steady state. This much is true, regardless of the nature of  $H$ , (whether for spin, or nuclear forces, etc.). In the special case of atomic systems when  $H = T + V$  where  $T$  is the kinetic energy operator, and  $V$  the potential, since  $\partial H / \partial \lambda = \partial V / \partial \lambda$  we can write

$$f'_\lambda = f_\lambda = -\frac{\partial U}{\partial \lambda} = -\int \psi^* \frac{\partial V}{\partial \lambda} \psi dv. \quad (3)$$

The actual calculation of forces in a real molecule by means of this theorem is not impractical. The  $\int \psi^* \psi (\partial V / \partial \lambda) dv$  is not too different from  $\int \psi^* \psi V dv$ , which must be calculated if the energy is to be found at all in the variational method. Although the theorem (3) is the most practical for actual calculations, it can be modified to get a clearer qualitative picture of what it means. Suppose, for example, the system for which  $\psi$  is the wave function contains several nuclei,  $\alpha$ , be  $X^\alpha$ ,  $Y^\alpha$ ,  $Z^\alpha$  or  $X_\mu^\alpha$  where  $\mu = 1, 2, 3$ , mean  $X$ ,  $Y$ ,  $Z$ . If we take our  $\lambda$  parameter to be one of these coordinates, the resultant force on the nucleus  $\alpha$  in the  $\mu$  direction will be given directly by

$$f_\mu^\alpha = -\int \psi \psi^* (\partial V / \partial X_\mu^\alpha) dv$$

from (3).

Now  $V$  is made up of three parts, the interaction of all nuclei with each other ( $V_{\alpha\beta}$ ), of each nucleus with an electron ( $V_{\beta i}$ ), and of each

electron with every other ( $V_{ij}$ ); or

$$V = \sum_{\alpha, \beta} V_{\alpha\beta} + \sum_{\beta, i} V_{\beta i} + \sum_{i, j} V_{ij}.$$

Suppose  $x_\mu^i$  are the coordinates of electron  $i$ , and as before  $X_\mu^\alpha$  those of nucleus  $\alpha$  of charge  $q_\alpha$ . Then  $V_{\beta i} = q_\beta e / R_{\beta i}$ , where

$$R_{\beta i}^2 = \sum_{\mu=1}^3 (X_\mu^\beta - x_\mu^i)^2.$$

So we see that

$$\frac{\partial V_{\beta i}}{\partial X_\mu^\alpha} = -\delta_{\beta\alpha} \frac{\partial V_{\beta i}}{\partial x_\mu^i} \quad \text{and that} \quad \frac{\partial V_{ij}}{\partial X_\mu^\alpha} = 0.$$

Then (3) leads to

$$\begin{aligned} f_\mu^\alpha &= + \int \psi \psi^* \sum_i \frac{\partial V_{\alpha i}}{\partial x_\mu^i} dv - \sum_\beta \int \frac{\partial V_{\alpha\beta}}{\partial X_\mu^\alpha} \psi \psi^* dv \\ &= \int \sum_i \frac{\partial V_{\alpha i}}{\partial x_\mu^i} \left[ \int i \int \psi \psi^* dv \right] dv - \sum_\beta \frac{\partial V_{\alpha\beta}}{\partial X_\mu^\alpha} \end{aligned} \quad (4)$$

since  $\partial V_{\alpha i} / \partial x_\mu^i$  does not involve any electron coordinate except those of electron  $i$ .  $\int i \int \psi \psi^* dv$  means the integral over the coordinates of all electrons except those of electron  $i$ . The last term has been reduced since  $\partial V_{\alpha\beta} / \partial X_\mu^\alpha$  does not involve the electron coordinates, and is constant as far as integration over these coordinates goes. This term gives ordinary Coulomb electrostatic repulsion between the nuclei and need not be considered further. Now  $e \int i \int \psi \psi^* dv$  is just the charge density distribution  $\rho_i(x)$  due to electron  $i$ , where  $e$  is the charge on one electron. The electric field  $E_\mu^\alpha(x^i)$  at any point  $x^i$  due to the nucleus  $\alpha$  is  $(1/e) \partial V_{\alpha i} / \partial x_\mu^i$ , so that (4) may be written

$$f_\mu^\alpha = \int \left[ \sum_i \rho_i(x) \right] E_\mu^\alpha(x) dv - \sum_\beta \frac{\partial V_{\alpha\beta}}{\partial X_\mu^\alpha}.$$

The  $3N$  space for  $N$  electrons has been reduced to a  $3$  space. This can be done since  $E_\mu^\alpha(x^i)$  depends only on  $x^i$  and is the same function of  $x^i$  no matter which  $i$  we pick. This implies the following conclusion:

The force on any nucleus (considered fixed) in any system of nuclei and electrons is just the classical electrostatic attraction exerted on the nucleus in question by the other nuclei and by

the electron charge density distribution for all electrons,

$$\rho(x) = \sum_i \rho_i(x).$$

It is possible to simplify this still further. Suppose we construct an electric field vector  $F$  such that

$$\nabla \cdot F = -4\pi\rho(x); \quad \nabla \times F = 0.$$

Now from the derivation of  $E_\mu^\alpha$  we know that it arises from the charge  $q_\alpha$  on nucleus  $\alpha$ , so that  $\nabla \cdot E^\alpha = 0$  except at the charge  $\alpha$  where its integral equals  $q_\alpha$ . Further,

$$-\sum_\beta \frac{\partial V_{\alpha\beta}}{\partial X_\mu^\alpha} = q_\alpha \left[ \sum_\beta E_\mu^\beta \right]_{at x^\alpha}.$$

Then

$$\begin{aligned} f_\mu^\alpha &= -\frac{1}{4\pi} \int (\nabla \cdot F) E_\mu^\alpha dv - \sum_\beta \frac{\partial V_{\alpha\beta}}{\partial X_\mu^\alpha} \\ &= +\frac{1}{4\pi} \int F_\mu (\nabla \cdot E_\mu^\alpha) dv - \sum_\beta \frac{\partial V_{\alpha\beta}}{\partial x_\mu^\alpha} \\ &= q_\alpha [F_\mu]_{at x^\alpha} + q_\alpha \left[ \sum_\beta E_\mu^\beta \right]_{at x^\alpha} \end{aligned} \quad (5)$$

the transformation of the integral being accomplished by integrating by parts. Or finally, the force on a nucleus is the charge on that nucleus times the electric field there due to all the electrons, plus the fields from the other nuclei. This field is calculated classically from the charge distribution of each electron and from the nuclei.

It now becomes quite clear why the strongest and most important attractive forces arise when there is a concentration of charge between two nuclei. The nuclei on each side of the concentrated charge are each strongly attracted to it. Thus they are, in effect, attracted toward each other. In a  $H_2$  molecule, for example, the anti-symmetrical wave function, because it must be zero exactly between the two  $H$  atoms, cannot concentrate charge between them. The symmetrical solution, however, can easily permit charge concentration between the nuclei, and hence it is only the solution which is symmetrical that leads to strong attraction, and the formation of a molecule, as is well known. It is

clearly seen that concentrations of charge between atoms lead to strong attractive forces, and hence, are properly called valence bonds.

Van der Waals' forces can also be interpreted as arising from charge distributions with higher concentration between the nuclei. The Schrödinger perturbation theory for two interacting atoms at a separation  $R$ , large compared to the radii of the atoms, leads to the result that the charge distribution of each is distorted from central symmetry, a dipole moment of order  $1/R^7$  being induced in each atom. The negative

charge distribution of each atom has its center of gravity moved slightly toward the other. It is not the interaction of these dipoles which leads to van der Waals' force, but rather the attraction of each nucleus for the distorted charge distribution of its *own* electrons that gives the attractive  $1/R^7$  force.

The author wishes to express his gratitude to Professor J. C. Slater who, by his advice and helpful suggestions, aided greatly in this work. He would also like to thank Dr. W. C. Herring for the latter's excellent criticisms.

## On Continued Gravitational Contraction

J. R. OPPENHEIMER AND H. SNYDER  
*University of California, Berkeley, California*

(Received July 10, 1939)

When all thermonuclear sources of energy are exhausted a sufficiently heavy star will collapse. Unless fission due to rotation, the radiation of mass, or the blowing off of mass by radiation, reduce the star's mass to the order of that of the sun, this contraction will continue indefinitely. In the present paper we study the solutions of the gravitational field equations which describe this process. In I, general and qualitative arguments are given on the behavior of the metrical tensor as the contraction progresses: the radius of the star approaches asymptotically its gravitational radius; light from the surface of the star is progressively reddened, and can escape over a progressively narrower range of angles. In II, an analytic solution of the field equations confirming these general arguments is obtained for the case that the pressure within the star can be neglected. The total time of collapse for an observer comoving with the stellar matter is finite, and for this idealized case and typical stellar masses, of the order of a day; an external observer sees the star asymptotically shrinking to its gravitational radius.

### I

RECENTLY it has been shown<sup>1</sup> that the general relativistic field equations do not possess any static solutions for a spherical distribution of cold neutrons if the total mass of the neutrons is greater than  $\sim 0.7\odot$ . It seems of interest to investigate the behavior of nonstatic solutions of the field equations.

In this work we will be concerned with stars which have large masses,  $>0.7\odot$ , and which have used up their nuclear sources of energy. A star under these circumstances would collapse under the influence of its gravitational field and release energy. This energy could be divided into four parts: (1) kinetic energy of motion of the

particles in the star, (2) radiation, (3) potential and kinetic energy of the outer layers of the star which could be blown away by the radiation, (4) rotational energy which could divide the star into two or more parts. If the mass of the original star were sufficiently small, or if enough of the star could be blown from the surface by radiation, or lost directly in radiation, or if the angular momentum of the star were great enough to split it into small fragments, then the remaining matter could form a stable static distribution, a white dwarf star. We consider the case where this cannot happen.

If then, for the late stages of contraction, we can neglect the gravitational effect of any escaping radiation or matter, and may still neglect the deviations from spherical symmetry

<sup>1</sup>J. R. Oppenheimer and G. M. Volkoff, Phys. Rev. 55, 374 (1939).

produced by rotation, the line element outside the boundary  $r_b$  of the stellar matter must take the form

$$ds^2 = e^\nu dt^2 - e^\lambda dr^2 - r^2(d\theta^2 + \sin^2 \theta d\varphi^2) \quad (1)$$

with

$$e^\nu = (1 - r_0/r)$$

and

$$e^\lambda = (1 - r_0/r)^{-1}.$$

Here  $r_0$  is the gravitational radius, connected with the gravitational mass  $m$  of the star by  $r_0 = 2mg/c^2$ , and constant. We should now expect that since the pressure of the stellar matter is insufficient to support it against its own gravitational attraction, the star will contract, and its boundary  $r_b$  will necessarily approach the gravitational radius  $r_0$ . Near the surface of the star, where the pressure must in any case be low, we should expect to have a local observer see matter falling inward with a velocity very close to that of light; to a distant observer this motion will be slowed up by a factor  $(1 - r_0/r_b)$ . All energy emitted outward from the surface of the star will be reduced very much in escaping, by the Doppler effect from the receding source, by the large gravitational red-shift,  $(1 - r_0/r_b)^{1/2}$ , and by the gravitational deflection of light which will prevent the escape of radiation except through a cone about the outward normal of progressively shrinking aperture as the star contracts. The star thus tends to close itself off from any communication with a distant observer; only its gravitational field persists. We shall see later that although it takes, from the point of view of a distant observer, an infinite time for this asymptotic isolation to be established, for an observer comoving with the stellar matter this time is finite and may be quite short.

Inside the star we shall still suppose that the matter is spherically distributed. We may then take the line element in the form (1). For this line element the field equations are

$$-8\pi T_1^1 = e^{-\lambda}(\nu'/r + 1/r^2) - 1/r^2, \quad (2)$$

$$8\pi T_4^4 = e^{-\lambda}(\lambda'/r - 1/r^2) + 1/r^2, \quad (3)$$

$$\begin{aligned} -8\pi T_2^2 &= -8\pi T_3^3 \\ &= e^{-\lambda} \left( \frac{\nu''}{2} + \frac{\nu'^2}{4} - \frac{\nu'\lambda'}{4} + \frac{\nu' - \lambda'}{2r} \right) \\ &\quad - e^{-\nu} (\ddot{\lambda}/2 + \dot{\lambda}^2/4 - \dot{\lambda}\dot{\nu}/4), \end{aligned} \quad (4)$$

$$8\pi T_4^1 = -8\pi e^{\nu-\lambda} T_1^4 = -e^{-\lambda} \dot{\lambda}/r; \quad (5)$$

in which primes represent differentiation with respect to  $r$  and dots differentiation with respect to  $t$ .

The energy-momentum tensor  $T_{\mu\nu}$  is composed of two parts: (1) a material part due to electrons, protons, neutrons and other nuclei, (2) radiation. The material part may be thought of as that of a fluid which is moving in a radial direction, and which in comoving coordinates would have a definite relation between the pressure, density, and temperature. The radiation may be considered to be in equilibrium with the matter at this temperature, except for a flow of radiation due to a temperature gradient.

We have been unable to integrate these equations except when we place the pressure equal to zero. However, one can obtain some information about the solutions from inequalities implied by the differential equations and from conditions for regularity of the solutions. From Eqs. (2) and (3) one can see that unless  $\lambda$  vanishes at least as rapidly as  $r^2$  when  $r \rightarrow 0$ ,  $T_4^4$  will become singular and that either or both  $T_1^1$  and  $\nu'$  will become singular. Physically such a singularity would mean that the expression used for the energy-momentum tensor does not take account of some essential physical fact which would really smooth the singularity out. Further, a star in its early stage of development would not possess a singular density or pressure; it is impossible for a singularity to develop in a finite time.

If, therefore,  $\lambda(r=0)=0$ , we can express  $\lambda$  in terms of  $T_4^4$ , for, integrating Eq. (3)

$$\lambda = -\ln \left\{ 1 - \frac{8\pi}{r} \int_0^r T_4^4 r^2 dr \right\}. \quad (6)$$

Therefore  $\lambda \geq 0$  for all  $r$  since  $T_4^4 \geq 0$ .

Now that we know  $\lambda \geq 0$ , it is easy to obtain some information about  $\nu'$  from Eq. (2);

$$\nu' \geq 0, \quad (7)$$

since  $\lambda$  and  $-T_1^1$  are equal to or greater than zero.

If we use clock time at  $r = \infty$ , we may take  $\nu(r = \infty) = 0$ . From this boundary condition and Eq. (7) we deduce

$$\nu \leq 0. \quad (8)$$

The condition that space be flat for large  $r$  is

$\lambda(r = \infty) = 0$ . Adding Eqs. (2) and (3) we obtain:

$$8\pi(T_4^4 - T_1^1) = e^\lambda(\lambda' + \nu')/r. \quad (9)$$

Since  $T_4^4$  is greater than zero and  $T_1^1$  is less than zero we conclude

$$\lambda' + \nu' \geq 0. \quad (10)$$

Because of the boundary conditions on  $\lambda$  and  $\nu$  we have

$$\lambda + \nu \leq 0. \quad (11)$$

For those parts of the star which are collapsing, i.e., all parts of the star except those being blown away by the radiation, Eq. (5) tells us that  $\dot{\lambda}$  is greater than zero. Since  $\lambda$  increases with time, it may (a) approach an asymptotic value uniformly as a function of  $r$ ; or (b) increase indefinitely, although certainly not uniformly as a function of  $r$ , since  $\lambda(r=0)=0$ . If  $\lambda$  were to approach a limiting value the star would be approaching a stationary state. However, we are supposing that the relationships between the  $T_{\nu}^{\mu}$  do not admit any stationary solutions, and therefore exclude this possibility. Under case (b) we might expect that for any value of  $r$  greater than zero,  $\lambda$  will become greater than any preassigned value if  $t$  is sufficiently large. If this were so the volume of the star

$$V = 4\pi \int_0^{r_b} e^{\lambda/2} r^2 dr \quad (12)$$

would increase indefinitely with time; since the mass is constant, the mean density in the star would tend to zero. We shall see, however, that for all values of  $r$  except  $r_0$ ,  $\lambda$  approaches a finite limiting value; only for  $r=r_0$  does it increase indefinitely.

## II

To investigate this question we will solve the field equations with the limiting form of the energy-momentum tensor in which the pressure is zero. When the pressure vanishes there are no static solutions to the field equations except when all components of  $T_{\nu}^{\mu}$  vanish. With  $p=0$  we have the free gravitational collapse of the matter. We believe that the general features of the solution obtained this way give a valid indication even for the case that the pressure is not zero, provided that the mass is great enough to cause collapse.

For the solution of this problem, we have found it convenient to follow the earlier work of Tolman<sup>2</sup> and use another system of coordinates, which are comoving with the matter. After finding a solution, we will introduce a coordinate transformation to put the line element in form (1).

We take a line element of the form:

$$ds^2 = d\tau^2 - e^{\bar{\omega}} dR^2 - e^{\omega}(d\theta^2 + \sin^2 \theta d\varphi^2). \quad (13)$$

Because the coordinates are comoving with the matter and the pressure is zero,

$$T_4^4 = \rho \quad (14)$$

and all other components of the energy momentum tensor vanish.

The field equations are:

$$8\pi T_1^1 = 0 = e^{-\omega} - e^{-\bar{\omega}} \frac{\omega'^2}{4} + \ddot{\omega} + \frac{3}{4}\dot{\omega}^2 = 0, \quad (15)$$

$$8\pi T_2^2 = 8\pi T_3^3 = 0 = -e^{-\bar{\omega}} \left( \frac{\omega''}{2} + \frac{\omega'^2}{4} - \frac{\dot{\omega}'\dot{\omega}'}{4} \right) + \frac{\ddot{\omega}}{2} + \frac{\dot{\omega}^2}{4} + \frac{\ddot{\omega}}{2} + \frac{\dot{\omega}^2}{4} + \frac{\dot{\omega}\dot{\omega}'}{4}, \quad (16)$$

$$8\pi T_4^4 = 8\pi\rho = e^{-\omega} - e^{-\bar{\omega}} \left( \omega'' + \frac{3}{4}\omega'^2 - \frac{\dot{\omega}'\dot{\omega}'}{2} \right) + \frac{\dot{\omega}^2}{4} + \frac{\dot{\omega}\dot{\omega}'}{2}, \quad (17)$$

$$8\pi e^{\bar{\omega}} T_4^1 = -8\pi T_1^4 = 0 = \frac{\omega'\dot{\omega}}{2} - \frac{\dot{\omega}\omega'}{2} + \dot{\omega}' \quad (18)$$

with primes and dots here and in the following representing differentiation with respect to  $R$  and  $\tau$ , respectively. The integral of Eq. (18) is given by Tolman:<sup>3</sup>

$$e^{\bar{\omega}} = e^{\omega} \omega'^2 / 4f^2(R) \quad (19)$$

with  $f^2(R)$  a positive but otherwise arbitrary function of  $R$ . We find a sufficiently wide class of solutions if we put  $f^2(R) = 1$ .

Substituting (19) in (15) with  $f^2(R) = 1$  we obtain

$$\ddot{\omega} + \frac{3}{4}\dot{\omega}^2 = 0. \quad (20)$$

<sup>2</sup> R. C. Tolman, Proc. Nat. Acad. Sci. **20**, 3 (1934).

<sup>3</sup> We wish to thank Professor R. C. Tolman and Mr. G. Omer for making this portion of the development available to us, and for helpful discussions.

The solution of this equation is:

$$e^\omega = (F\tau + G)^{4/3}, \quad (21)$$

in which  $F$  and  $G$  are arbitrary functions of  $R$ .

The substitution of (19) in (16) gives a result equivalent to (20). Therefore the solution of the field equations is (21).

For the density we obtain from (17), (19), and (21)

$$8\pi\rho = 4/3(\tau + G/F)^{-1}(\tau + G'/F')^{-1}. \quad (22)$$

There is less real freedom in (21) than is apparent from the two arbitrary functions  $F$  and  $G$ ; for taking  $R$  a function of a new variable  $R^*$  the differential equations (15), (17) and (18) will remain of the same form. We may therefore choose

$$G = R^{4/3}. \quad (23)$$

At a particular time, say  $\tau$  equal zero, we may assign the density as a function of  $R$ . Eq. (22) then becomes a first-order differential equation for  $F$ .

$$FF' = 9\pi R^2 \rho_0(R). \quad (24)$$

The solution of this equation contains only one arbitrary constant. We now see that the effect of setting  $f^2(R)$  equal to one allows us to assign only a one-parameter family of functions for the initial values of  $\rho_0$ , whereas in general one should be able to assign the initial values of  $\rho_0$  arbitrarily.

We now take, as a particular case of (24):

$$FF' = \begin{cases} \text{const.} \times R^2; & \text{const.} > 0; \\ 0 & R < R_b \\ & ; R > R_b. \end{cases} \quad (25)$$

A particular solution of this equation is:

$$F = \begin{cases} -\frac{3}{2}r_0^{\frac{1}{3}}(R/R_b)^{\frac{2}{3}}; & R < R_b \\ -\frac{3}{2}r_0^{\frac{1}{3}} & ; R > R_b \end{cases} \quad (26)$$

in which the constant  $r_0$  is introduced for convenience, and is the gravitational radius of the star.

We wish to find a coordinate transformation which will change the line element into form (1). It is clear, by comparison of (1) and (13), that we must take

$$e^{\omega/2} = (F\tau + G)^{\frac{2}{3}} = r. \quad (27)$$

A new variable  $t$  which is a function of  $\tau$  and  $R$  must be introduced so that the  $g_{\mu\nu}$  are of the

same form as those in Eq. (1). Using the contravariant form of the metric tensor, we find that:

$$g^{44} = e^{-\nu} = t^2 - t'^2/r'^2 = t^2(1 - \dot{r}^2), \quad (28)$$

$$g^{11} = -e^{-\lambda} = -(1 - \dot{r}^2), \quad (29)$$

$$g^{14} = 0 = \dot{t}\dot{r} - t'/r'. \quad (30)$$

Here (30) is a first-order partial differential equation for  $t$ . Using the values of  $r$  given by (27), and the values of  $F$  and  $G$  given by (26) and (23) we find:

$$t'/\dot{r} = \dot{r}r' = \begin{cases} -(r_0R)^{\frac{1}{3}}[R^{\frac{2}{3}} - \frac{3}{2}r_0^{\frac{1}{3}}\tau]^{-\frac{2}{3}}; & R > R_b \\ -r_0^{\frac{1}{3}}RR_b^{-\frac{2}{3}}[1 - \frac{3}{2}r_0^{\frac{1}{3}}\tau R_b^{-\frac{2}{3}}]^{\frac{1}{3}}; & R < R_b. \end{cases} \quad (31)$$

The general solution of (31) is:

$$\begin{aligned} t = L(x) \text{ for } R > R_b, \text{ with } x = & \frac{2}{3r_0^{\frac{1}{3}}}(R^{\frac{2}{3}} - r^{\frac{2}{3}}) \\ & - 2(rr_0)^{\frac{1}{3}} + r_0 \ln \frac{r^{\frac{1}{3}} + r_0^{\frac{1}{3}}}{r^{\frac{1}{3}} - r_0^{\frac{1}{3}}} \quad (32) \\ t = M(y) \text{ for } R < R_b, \text{ with } y = & \frac{1}{2}[(R/R_b)^2 - 1] \\ & + R_b r / r_0 R, \end{aligned}$$

where  $L$  and  $M$  are completely arbitrary functions of their arguments.

Outside the star, where  $R$  is greater than  $R_b$ , we wish the line element to be of the Schwartzchild form, since we are again neglecting the gravitational effect of any escaping radiation; thus

$$e^\lambda = (1 - r_0/r)^{-1} \quad (33)$$

$$e^\nu = (1 - r_0/r). \quad (34)$$

This requirement fixes the form of  $L$ ; from (28) we can show that we must take  $L(x) = x$ , or

$$t = x. \quad (35)$$

At the surface of the star,  $R$  equal  $R_b$ , we must have  $L$  equal to  $M$  for all  $\tau$ . The form of  $M$  is determined by this condition to be:

$$\begin{aligned} t = M(y) = & \frac{2}{3}r_0^{-\frac{1}{3}}(R_b^{\frac{2}{3}} - r_0^{\frac{1}{3}}y^{\frac{2}{3}}) \\ & - 2r_0y^{\frac{1}{3}} + r_0 \ln \frac{y^{\frac{1}{3}} + 1}{y^{\frac{1}{3}} - 1}. \quad (36) \end{aligned}$$

Eq. (36), together with (27), defines the transformation from  $R$ ,  $\tau$  to  $r$  and  $t$ , and implicitly, from (28) and (29), the metrical tensor.

We now wish to find the asymptotic behavior of  $e^\lambda$ ,  $e^\nu$ , and  $\tau$  for large values of  $t$ . When  $t$  is large we obtain the approximate relation from Eqs. (36) and (27):

$$\begin{aligned} t \sim -r_0 \ln & \left\{ \frac{1}{2}[(R/R_b)^2 - 3] \right. \\ & \left. + R_b/r_0(1 - 3r_0^{\frac{1}{2}}\tau/2R_b^2)^{\frac{1}{2}} \right\}. \quad (37) \end{aligned}$$

From this relation we see that for a fixed value of  $R$  as  $t$  tends toward infinity,  $\tau$  tends to a finite limit, which increases with  $R$ . After this time  $r_0$  an observer comoving with the matter would not be able to send a light signal from the star; the cone within which a signal can escape has closed entirely. For a star which has an initial density of one gram per cubic centimeter and a mass of  $10^{33}$  grams this time  $r_0$  is about a day.

Substituting (27) and (37) into (28) and (29) we find

$$e^{-\lambda} \simeq 1 - (R/R_b)^2 \{e^{-t/r_0} + \frac{1}{2}[3 - (R/R_b)^2]\}^{-1}, \quad (38)$$

$$e^\nu \simeq e^{\lambda - 2t/r_0} \{e^{-t/r_0} + \frac{1}{2}[3 - (R/R_b)^2]\}. \quad (39)$$

For  $R$  less than  $R_b$ ,  $e^\lambda$  tends to a finite limit as  $t$  tends to infinity. For  $R$  equal to  $R_b$ ,  $e^\lambda$  tends to infinity like  $e^{t/r_0}$  as  $t$  approaches infinity. Where  $R$  is less than  $R_b$ ,  $e^\nu$  tends to zero like  $e^{-2t/r_0}$  and where  $R$  is equal to  $R_b$ ,  $e^\nu$  tends to zero like  $e^{-t/r_0}$ .

This quantitative account of the behavior of  $e^\lambda$  and  $e^\nu$  can supplement the qualitative discussion given in I. For  $\lambda$  tends to a finite limit for  $r < r_0$  as  $t$  approaches infinity, and for  $r = r_0$  tends to infinity. Also for  $r \leq r_0$ ,  $\nu$  tends to minus infinity. We expect that this behavior will be realized by all collapsing stars which cannot end in a stable stationary state. Of course, actual stars would collapse more slowly than the example which we studied analytically because of the effect of the pressure of matter, of radiation, and of rotation.



## Letters to the Editor

*Prompt publication of brief reports of important discoveries in physics may be secured by addressing them to this department. Closing dates for this department are, for the first issue of the month, the eighteenth of the preceding month, or the second issue, the third of the month. Because of the late closing dates for the section no proof can be shown to authors. The Board of Editors does not hold itself responsible for the opinions expressed by the correspondents. Communications should not in general exceed 600 words in length.*

**Communications should not in general exceed 600 words in length.**

D. W. KERST

University of Illinois, Urbana, Illinois

(Received October 15, 1940)

## Acceleration of Electrons by Magnetic Induction

For some time it has been realized that it might be possible to make use of the electromotive force induced by a changing magnetic flux to accelerate charged particles traveling in an orbit around the changing flux. Although previous attempts to accelerate electrons by this means have been unsuccessful<sup>1</sup> <sup>2</sup> careful examination showed that it should be possible to get good magnetic focusing by the proper arrangement of a magnetic field to guide the electrons around the changing flux and that if the rate of change of flux within the orbit were sufficiently high it would be possible to capture electrons in usable orbits and that vacuum requirements should not be difficult to satisfy.

It seemed feasible to attempt the experiment with a 600-cycle per second magnetic field, since a sufficiently high rate of change of flux would be obtained and since it seemed that it would not be necessary to have a vacuum better than  $10^{-6}$  millimeter of mercury in the acceleration chamber, in spite

---

<sup>1</sup>Wideröe, Archiv f. Elektrotechnik **21**, 400 (1938).

<sup>2</sup>E. T. S. Walton. Proc. Camb. Phil. Soc. **25**, 469 (1929).

of the fact that at this frequency the length of the electron path would be of the order of  $10^7$  centimeters.

To hold the electrons in the acceleration chamber for such a long path it is necessary to fulfill the condition that  $\phi = 2\pi R_0^2 H$ , where  $\phi$  is the flux enclosed by the orbit and  $H$  is the magnetic field at the orbit which causes the electrons to travel in a circle of radius  $R_0$ . When this condition is satisfied, the electron orbit neither shrinks nor expands, and the electrons can be accelerated by increasing  $\phi$  and  $H$  together.

A laminated electromagnet with pole faces 8 inches in diameter, which satisfied all the necessary conditions, was constructed. The stable orbit was shrunk from  $R_0$  toward the position of a tungsten target by causing saturation of the portion of the magnetic circuit which supplied the flux through the center of the orbit. *X*-rays produced by the impact of the electrons upon the target showed that the accelerator operated, and a lead collimator in front of a Geiger-Müller counter showed that the only portion of the acceleration chamber from which *x*-rays came was the target.

By taking the sweep voltage for an oscillosograph from a coil surrounding the core of the magnet and putting the pulses from the Geiger-Müller counter circuit on the vertical deflection plates, the phase of the magnetic field at which the electrons struck the target could be determined. It was possible to hold the electrons in the acceleration chamber for one-fourth of a cycle during which the magnetic field changed from a low value to its maximum. Conservative estimates of the magnetic field at the target when the electrons strike it indicated that the energy of the electrons was about 2.2 Mev. This estimate was substantiated by a comparison of the absorption of the *x*-rays in lead with published data on the absorption of *x*-rays produced by 2-million-volt electrons<sup>3</sup>. After filtering the *x*-rays from the accelerator through about 1.8 cm of lead, their absorption coefficient is  $0.57 \text{ cm}^{-1}$ . A correction had to be made for scattering of *x*-rays from the magnet yoke. Since the absorption coefficient for *x*-rays produced by 2.0-Mev electrons is  $0.62 \text{ cm}^{-1}$ , the electrons in the new accelerator must have reached about 2.35-Mev energy before striking the target. The absorption measurements were taken with Lauritsen electroscopes, and calibration of the electroscopes showed that the intensity of the radiation was greater than the intensity of the gamma-rays from TO millicuries of radium.

Of several suggestions which have been made for naming the apparatus, induction accelerator seems to be the shortest descriptive one.

---

<sup>3</sup>D. L. Northrup and L. C. Van Atta, Am. J. Roentgenology and Radium Therapy **41**, 633 (1939).

It has been a great help to be able to discuss the theoretical aspects of the accelerator with Professor R. Serber and Professor H. M. Mott-Smith.

D. W. KERST



The Connection Between Spin and Statistics<sup>1</sup>

W. Pauli  
Princeton, New Jersey  
(Received August 19, 1940)

— — ◇ ◇ — —  
Reprinted in "Quantum Electrodynamics", edited by Julian Schwinger  
— — ◇ ◇ — —

**Abstract**

In the following paper we conclude for the relativistically invariant wave equation for free particles: From postulate (I), according to which the energy must be positive, the necessity of *Fermi-Dirac* statistics for particles with arbitrary half-integral spin; from postulate (II), according to which observables on different space-time points with a space-like distance are commutable, the necessity of *Einstein-Base* statistics for particles with arbitrary integral spin. It has been found useful to divide the quantities which are irreducible against Lorentz transformations into four symmetry classes which have a commutable multiplication like  $+1, -1, +\epsilon, -\epsilon$  with  $\epsilon^2 = 1$ .

<sup>1</sup>This paper is part of a report which was prepared by the author for the Solvay Congress 1939 and in which slight improvements have since been made. In view of the unfavorable times, the Congress did not take place, and the publication of the reports has been postponed for an indefinite length of time. The relation between the present discussion of the connection between spin and statistics, and the somewhat less general one of Belinfante, based on the concert of charge invariance, has been cleared up by W. Pauli and J. Belinfante, *Physica* 7, 177 (1940).

## § 1. UNITS AND NOTATIONS

Since the requirements of the relativity theory and the quantum theory are fundamental for every theory, it is natural to use as units the vacuum velocity of light  $c$ , and Planck's constant divided by  $2\pi$  which we shall simply denote by  $\hbar$ . This convention means that all quantities are brought to the dimension of the power of a length by multiplication with powers of  $\hbar$  and  $c$ . The reciprocal length corresponding to the rest mass  $m$  is denoted by  $\kappa = mc/\hbar$ .

As time coordinate we use accordingly the length of the light path. In specific cases, however, we do not wish to give up the use of the imaginary time coordinate. Accordingly, a tensor index denoted by small Latin letters  $i$ , refers to the imaginary time coordinate and runs from 1 to 4. A special convention for denoting the complex conjugate seems desirable. Whereas for quantities with the index 0 an asterisk signifies the complex-conjugate in the ordinary sense (e.g., for the current vector  $S_i$  the quantity  $S_0^*$  is the complex conjugate of the charge density  $S_0$ ), in general  $U_{ik\dots}^*$  signifies: the complex-conjugate of  $U_{ik\dots}$  multiplied with  $(-1)^n$ , where  $n$  is the number of occurrences of the digit 4 among the  $i, k, \dots$  (e.g.  $S_4 = iS_0$ ,  $S_4^* = iS_0^*$ ).

Dirac's spinors  $u_\rho$ , with  $\rho = 1, \dots, 4$  have always a Greek index running from 1 to 4, and  $u_\rho^*$  means the complex-conjugate of  $u_\rho$ , in the ordinary sense.

Wave functions, insofar as they are ordinary vectors or tensors, are denoted in general with capital letters,  $U_i, U_{ik\dots}$ . The symmetry character of these tensors must in general be added explicitly. As classical fields the electromagnetic and the gravitational fields, as well as fields with rest mass zero, take a special place, and are therefore denoted with the usual letters  $\varphi_i$ ,  $f_{ik} = -f_{ki}$  and  $g_{ik} = g_{ki}$  respectively.

The energy-momentum tensor  $T_{ik}$ , is so defined, that the energy-density  $W$  and the momentum density  $G_\kappa$  are given in natural units by  $W = -T_{44}$  and  $G_\kappa = -iT_{\kappa 4}$  with  $k = 1, 2, 3$ .

## § 2. IRREDUCIBLE TENSORS. DEFINITION OF SPINS

We shall use only a few general properties of those quantities which transform according to irreducible representations of the Lorentz group.<sup>2</sup> The

---

<sup>2</sup>See B. L. v. d. Waerden, *Die gruppentheoretische Methode in der Quantentheorie* (Berlin, 1932).

proper Lorentz group is that continuous linear group the transformations of which leave the form

$$\sum_{k=1}^4 x_k^2 = \mathbf{x}^2 - x_0^2$$

invariant and in addition to that satisfy the condition that they have the determinant +1 and do not reverse the time. A tensor or spinor which transforms irreducibly under this group can be characterized by two integral positive numbers  $(p, q)$ . (The corresponding “angular momentum quantum numbers”  $(j, k)$  are then given by  $p = 2j + 1$ ,  $q = 2k + 1$ , with integral or half-integral  $j$  and  $k$ ).<sup>3</sup> The quantity  $U(j, k)$  characterized by  $(j, k)$  has  $p \cdot q = (2j+1)(2k+1)$  independent components. Hence to  $(0, 0)$  corresponds the scalar, to  $(\frac{1}{2}, \frac{1}{2})$  the vector, to  $(1, 0)$  the self-dual skew-symmetrical tensor, to  $(1, 1)$  the symmetrical tensor with vanishing spur, etc. Dirac’s spinor it, reduces to two irreducible quantities  $(\frac{1}{2}, 0)$  and  $(0, \frac{1}{2})$  each of which consists of two components. If  $U(j, k)$  transforms according to the representation

$$U'_r = \sum_{s=1}^{(2j+1)(2k+1)} \Lambda_{rs} U_s,$$

then  $U^*(k, j)$  transforms according to the complex-conjugate representation  $\Lambda^*$ . Thus for  $k = j$ ,  $\Lambda^* = \Lambda$ . This is true only if the components of  $U(j, k)$  and  $U(k, j)$  are suitably ordered. For an arbitrary choice of the components, a similarity transformation of  $\Lambda$  and  $\Lambda^*$  would have to be added. In view of § 1 we represent generally with  $U^*$  the quantity the transformation of which is equivalent to  $\Lambda^*$  if the transformation of  $U$  is equivalent to  $\Lambda$ .

The most important operation is the reduction of the product of two quantities

$$U_1(j_1, k_1) \cdot U_2(j_2, k_2)$$

which, according to the well-known rule of the composition of angular momenta, decompose into several  $U(j, k)$  where, independently of each other  $j, k$  run through the values

$$j = j_1 + j_2, j_1 + j_2 - 1, \dots, |j_1 - j_2|$$

$$k = k_1 + k_2, k_1 + k_2 - 1, \dots, |k_1 - k_2|.$$

By limiting the transformations to the subgroup of space rotations alone, the distinction between the two numbers  $j$  and  $k$  disappears and  $U(j, k)$

---

<sup>3</sup>In the spinor calculus this is a spinor with  $2j$  undotted and  $2k$  dotted indices.

behaves under this group just like the product of two irreducible quantities  $U(j)U(k)$  which in turn reduces into several irreducible  $U(l)$  each having  $2l+1$  components, with

$$l = j + k, j + k - 1, \dots, |j - k|.$$

Under the space rotations the  $U(l)$  with integral  $l$  transform according to single-valued representation, whereas those with half-integral  $l$  transform according to double-valued representations. Thus the unreduced quantities  $T(j, k)$  with integral (half-integral)  $j + k$  are single-valued (double-valued).

If we now want to determine the spin value of the particles which belong to a given field it seems at first that these are given by  $l = j + k$ . Such a definition would, however, not correspond to the physical facts, for there then exists no relation of the spin value with the number of independent plane waves, which are possible in the absence of interaction) for given values of the components  $k$  in the phase factor  $\exp i(\mathbf{k}\mathbf{x})$ . In order to define the spin in an appropriate fashion,<sup>4</sup> we want to consider first the case in which the rest mass  $m$  of all the particles is different from zero. In this case we make a transformation to the rest system of the particle, where all the space components of  $k_i$ , are zero, and the wave function depends only on the time. In this system we reduce the field components, which according to the field equations do not necessarily vanish, into parts irreducible against space rotations. To each such part, with  $r = 2s+1$  components<sup>i</sup> belong  $r$  different eigenfunctions which under space rotations transform among themselves and which belong to a particle with spin  $s$ . If the field equations describe particles with only one spin value there then exists in the rest system only one such irreducible group of components. From the Lorentz invariance, it follows, for an arbitrary system of reference, that  $r$  or  $\sum r$  eigenfunctions always belong to a given arbitrary  $k_i$ . The number of quantities  $U(j, k)$  which enter the theory is, however, in a general coordinate system more complicated, since these quantities together with the vector  $k_i$  have to satisfy several conditions.

In the case of zero rest mass there is a special degeneracy because, as has been shown by Fierz, this case permits a gauge transformation of the second kind.<sup>5</sup> If the field now describes only one kind of particle with the rest mass zero and a certain spin value, then there are for a given value of  $k_i$  only two states, which cannot be transformed into each other by a gauge

---

<sup>4</sup>see M. Fierz, Helv. Phys. acta **12**, 3 (1939); also L. de Broglie, Comptes rendus **208**, 1697 (1939); **209**, 265 (1939).

<sup>5</sup>By “gauge-transformation of the first kind” we understand a transformation  $U \rightarrow U e^{i\alpha}$   $U^* \rightarrow U^* e^{-i\alpha}$  with an arbitrary space and time function  $\alpha$ . By “gauge-transformation of

transformation. The definition of spin may, in this case, not be determined so far as the physical point of view is concerned because the total angular momentum of the field cannot be divided up into orbital and spin angular momentum by measurements. But it is possible to use the following property for a definition of the spin. If we consider, in the  $q$  number theory, states where only one particle is present, then not all the eigenvalues  $j(j+1)$  of the square of the angular momentum are possible. But  $j$  begins with a certain minimum value  $s$  and takes then the values  $s, s+1, \dots$ <sup>6</sup> This is only the case for  $m=0$ . For photons,  $s=1$ ,  $j=0$  is not possible for one single photon.<sup>7</sup> For gravitational quanta  $s=l$  and the values  $j=0$  and  $j=1$  do not occur.

In an arbitrary system of reference and for arbitrary rest masses, the quantities  $U$  all of which transform according to double-valued (single-valued) representations with half-integral (integral)  $j+k$  describe only particles with half-integral (integral) spin. A special investigation is required only when it is necessary to decide whether the theory describes particles with one single spin value or with several spin values.

### § 3. PROOF OF THE INDEFINITE CHARACTER OF THE CHARGE IN CASE OF INTEGRAL AND OF THE ENERGY IN CASE OF HALF-INTEGRAL SPIN

We consider first a theory which contains only  $U$  with integral  $j+k$ , i.e., which describes particles with integral spins only. It is not assumed that only particles with one single spin value will be described, but all particles shall have integral spin.

We divide the quantities  $U$  into two classes: (1) the “+1 class” with  $j$  integral,  $k$  integral; (2) the “−1 class” with  $j$  half-integral,  $k$  half-integral.

The notation is justified because, according to the indicated rules about the reduction of a product into the irreducible constituents under the Lorentz

---

the second kind” we understand a transformation of the type

$$\varphi_k \rightarrow \varphi_k - \frac{1}{\epsilon} i \frac{\partial \alpha}{\partial x_k}$$

as for those of the electromagnetic potentials.

<sup>6</sup>The general proof for this has been given by M. Fierz, Helv. Phys. Acta 13, 45 (1940).

<sup>7</sup>See for instance W. Pauli in the article “Wellen-mechanik” in the Handbuch der Physik, Vol. **24/2**, p. 260.

group, the product of two quantities of the +1 class or two quantities of the -1 class contains only quantities of the +1 class, whereas the product of a quantity of the +1 class with a quantity of the -1 class contains only quantities of the -1 class. It is important that the complex conjugate  $U^*$  for which  $j$  and  $k$  are interchanged belong to the same class as  $U$ . As can be seen easily from the multiplication rule, tensors with even (odd) number of indices reduce only to quantities of the +1 class (-1 class). The propagation vector  $k_i$  we consider as belonging to the -1 class, since it behaves after multiplication with other quantities like a quantity of the -1 class.

We consider now a homogeneous and linear equation in the quantities  $U$  which, however, does not necessarily have to be of the first order. Assuming a plane wave, we may put  $k_i$  for  $-i\partial/\partial x_l$ . Solely on account of the invariance against the *proper* Lorentz group it must be of the typical form

$$\sum kU^+ = \sum U^-, \quad \sum kU^- = \sum U^+. \quad (1)$$

This typical form shall mean that there may be as many different terms of the same type present, as there are quantities  $U^*$  and  $U^-$ . Furthermore, among the  $U^*$  may occur the  $U^+$  as well as the  $(U^+)^*$ , whereas other  $U$  may satisfy reality conditions  $U = U^*$ . Finally we have omitted an even number of  $k$  factors. These may be present in arbitrary number in the term of the sum on the left- or right-hand side of these equations. It is now evident that these equations remain invariant under the substitution

$$\begin{aligned} k_i &\rightarrow -k_i; & U^+ &\rightarrow U^+, \quad [(U^+) \rightarrow (U^+)^*]; \\ U^- &\rightarrow -U^-, \quad [(U^-)^* \rightarrow -(U^-)^* \rightarrow -(U^-)^*]. \end{aligned} \quad (2)$$

Let us consider now tensors  $T$  of even rank (scalars, skew-symmetrical or symmetrical tensors of the 2nd rank, etc.), which are composed quadratically or bilinearly of the  $U'$ s. They are then composed solely of quantities with even  $j$  and even  $k$  and thus are of the typical form

$$T \sim \sum U^+ U^+ + \sum U^- U^- + \sum U^+ k U^-, \quad (3)$$

where again a possible even number of  $k$  factors is omitted and no distinction between  $U$  and  $U^*$  is made. Under the substitution (2) they remain unchanged,  $T \rightarrow T$ .

The situation is different for tensors of odd rank  $S$  (vectors, etc.) which consist of quantities with half-integral  $j$  and half-integral  $k$ . These are of the typical form

$$S \sim \sum U^+ k U^+ + \sum U^- k U^- + \sum U^- \quad (4)$$

and hence change the sign under the substitution (2),  $S \rightarrow -S$ . Particularly is this the case for the current vector  $s_i$ . To the transformation  $k_i \rightarrow -k_i$ , belongs for arbitrary wave packets the transformation  $x_i \rightarrow -x_i$ , and it is remarkable that from the invariance of Eq. (I) against the proper Lorentz group alone there follows an invariance property for the change of sign of all the coordinates. In particular, the indefinite character of the current density and the total charge for even spin follows, since to every solution of the field equations belongs another solution for which the components of  $s_k$ , change their sign. The definition of a definite particle density for even spin which transforms like the 4-component of a vector is therefore impossible.

We now proceed to a discussion of the somewhat less simple case of half-integral spins. Here we divide the quantities  $U$ , which have half-integral  $j + k$ , in the following fashion: (3) the “ $+\epsilon$  class” with  $j$  integral  $k$  half-integral, (4) the “ $-\epsilon$  class” with  $j$  half-integral  $k$  integral.

The multiplication of the classes (1), ..., (4), follows from the rule  $\epsilon^2 = 1$  and the commutability of the multiplication. This law remains unchanged if  $\epsilon$  is replaced by  $-\epsilon$ .

We can summarize the multiplication law between the different classes in the following multiplication table:

	1	-1	$\epsilon$	$-\epsilon$
1	1	-1	$\epsilon$	$-\epsilon$
-1	-1	+1	$-\epsilon$	$+\epsilon$
$\epsilon$	$-\epsilon$	$-\epsilon$	+1	-1
$-\epsilon$	$-\epsilon$	$\epsilon$	-1	+1

We notice that these classes have the multiplication law of Klein's “four-group.”

It is important that here the complex-conjugate quantities for which  $j$  and  $k$  are interchanged do not belong to the same class, so that

$$\begin{array}{ll} U^{+\epsilon}, (U^{-\epsilon})^* & \text{belong to the } +\epsilon \text{ class} \\ U^{-\epsilon}, (U^{+\epsilon})^* & \text{belong to the } -\epsilon \text{ class.} \end{array}$$

We shall therefore cite the complex-conjugate quantities explicitly. (One could even choose the  $U^{+\epsilon}$  suitably so that *all* quantities of the  $-\epsilon$  class are of the form  $(U^{+\epsilon})^*$ ).

Instead of (1) we obtain now as typical form

$$\begin{aligned} \sum kU^{+\epsilon} + \sum k(U^{-\epsilon})^* &= \sum U^{-\epsilon} + \sum (U^{+\epsilon})^* \\ \sum kU^{-\epsilon} + \sum k(U^{+\epsilon})^* &= \sum U^{+\epsilon} + \sum (U^{-\epsilon})^*, \end{aligned} \tag{5}$$

since a factor  $k$  or  $-i\partial/\partial x$  always changes the expression from one of the classes  $+\epsilon$  or  $-\epsilon$  into the other. As above, an even number of  $k$  factors have been omitted.

Now we consider instead of (2) the substitution

$$\begin{aligned} k_i &\rightarrow -k_i; \quad U^{+\epsilon} \rightarrow iU^{+\epsilon}; \quad (U^{-\epsilon})^* \rightarrow i(U^{-\epsilon})^*; \\ (U^{+\epsilon})^* &\rightarrow -i(U^{+\epsilon})^*; \quad U^{-\epsilon} \rightarrow -iU^{-\epsilon}. \end{aligned} \quad (6)$$

This is in accord with the algebraic requirement of the passing over to the complex conjugate, as well as with the requirement that quantities of the same class as  $U^{+\epsilon}$ ,  $(U^{-\epsilon})^*$  transform in the same way. Furthermore, it does not interfere with possible reality conditions of the type  $U^{+\epsilon} = (U^{-\epsilon})^*$  or  $U^{-\epsilon} = (U^{+\epsilon})^*$ . Equations (5) remain unchanged under the substitution (6).

We consider again tensors of even rank (scalars, tensors of 2nd rank, etc.), which are composed bilinearly or quadratically of the  $U$  and their complex-conjugate. For reasons similar to the above they must be of the form

$$\begin{aligned} T \sim & \sum U^{+\epsilon}U^{+\epsilon} + \sum U^{-\epsilon}U^{-\epsilon} + \sum U^{+\epsilon}kU^{-\epsilon} + \sum U^{+\epsilon}(U^{-\epsilon})^* \\ & + \sum U^{-\epsilon}(U^{+\epsilon})^* + \sum (U^{-\epsilon})^*kU^{-\epsilon} + \sum (U^{+\epsilon})^*kU^{+\epsilon} + \sum (U^{-\epsilon})^*k(U^{+\epsilon})^* \\ & \sum (U^{-\epsilon})^*(U^{-\epsilon})^* + \sum (U^{+\epsilon})^*(U^{+\epsilon})^*. \end{aligned} \quad (7)$$

Furthermore, the tensors of odd rank (vectors, etc.) must be of the form

$$\begin{aligned} S \sim & \sum U^{+\epsilon}kU^{+\epsilon} + \sum U^{-\epsilon}kU^{-\epsilon} + \sum U^{+\epsilon}U^{-\epsilon} + \sum U^{+\epsilon}k(U^{-\epsilon})^* \\ & + \sum U^{-\epsilon}k(U^{+\epsilon})^* + \sum U^{-\epsilon}(U^{-\epsilon})^* + \sum U^{+\epsilon}(U^{+\epsilon})^* + \sum (U^{-\epsilon})^*k(U^{-\epsilon})^* + \\ & \sum (U^{+\epsilon})^*k(U^{+\epsilon})^* + \sum (U^{-\epsilon})^*(U^{+\epsilon})^*. \end{aligned} \quad (8)$$

*The result of the substitution (6) is now the opposite of the result of the substitution (2): the tensors of even rank change their sign, the tensors of odd rank remain unchanged:*

$$T \rightarrow -T; \quad S \rightarrow +S. \quad (9)$$

In case of half-integral spin, therefore, a positive definite energy density, as well as a positive definite total energy, is impossible. The latter follows from the fact, that, under the above substitution, the energy density in every space-time point changes its sign as a result of which the total energy changes also its sign.

It may be emphasized that it was not only unnecessary to assume that the wave equation is of the first order,<sup>8</sup> but also that the question is left

---

<sup>8</sup>But we exclude operation like  $(k^2 + \kappa^2)^{1/2}$ , which operate at finite distances in the coordinate space.

open whether the theory is also invariant with respect to space reflections ( $\mathbf{x}' = -\mathbf{x}$ ,  $x'_0 = x_0$ ). This scheme covers therefore also Dirac's two component wave equations (with rest mass zero).

These considerations do not prove that for integral spins there always exists a definite energy density and for half-integral spins a definite charge density. In fact, it has been shown by Fierz<sup>9</sup> that this is not the case for spin  $> 1$  for the densities. There exists, however (in the  $c$  number theory), a definite total charge for half-integral spins and a definite total energy for the integral spins. The spin value  $\frac{1}{2}$  is discriminated through the possibility of a definite charge density, and the spin values 0 and 1 are discriminated through the possibility of defining a definite energy density. Nevertheless, the present theory permits arbitrary values of the spin quantum numbers of elementary particles as well as arbitrary values of the rest mass, the electric charge, and the magnetic moments of the particles.

#### **§ 4. QUANTIZATION OF THE FIELDS IN THE ABSENCE OF INTERACTIONS. CONNECTION BETWEEN SPIN AND STATISTICS**

The impossibility of defining in a physically satisfactory way the particle density in the case of integral spin and the energy density in the case of half-integral spins in the  $c$ -number theory is an indication that a satisfactory interpretation of the theory within the limits of the one-body problem is not possible.<sup>10</sup> In fact, all relativistically invariant theories lead to particles, which in external fields can be emitted and absorbed in pairs of opposite charge for electrical particles and singly for neutral particles. The fields must, therefore, undergo a second quantization. For this we do not wish to apply here the canonical formalism, in which time is unnecessarily sharply distinguished from space, and which is only suitable if there are no supplementary conditions between the canonical variables.<sup>11</sup> Instead, we shall apply here a generalization of this method which was applied for the

---

<sup>9</sup>M. Fierz, Helv. Phys. Acta 12, 3 (1939).

<sup>10</sup>The author therefore considers as not conclusive the original argument of Dirac, according to which the field equation must be of the first order.

<sup>11</sup>On account of the existence of such conditions the canonical formalism is not applicable for spin  $> 1$  and therefore the discussion about the connection between spin and statistics by J. S. de Wet, Phys. Rev. **57**, 646 (1940), which is based on that formalism is not general enough.

first time by Jordan and Pauli to the electromagnetic field.<sup>12</sup> This method is especially convenient in the absence of interaction, where all fields  $U^{(r)}$  satisfy the wave equation of the second order

$$\square U^{(r)} - \kappa^2 U^{(r)} = 0,$$

where

$$\square \equiv \sum_{k=1}^4 \frac{\partial^2}{\partial x_k^2} = \Delta - \frac{\partial^2}{\partial x_0^2}$$

and  $\kappa$  is the rest mass of the particles in units  $hbar/c$ .

An important tool for the second quantization is the invariant  $D$  function, which satisfies the wave equation (9) and is given in a periodicity volume  $V$  of the eigenfunctions by

$$D(\mathbf{x}, x_0) = \frac{1}{V} \sum \exp[i(\mathbf{k}\mathbf{x})] \frac{\sin k_0 x_0}{k_0} \quad (10)$$

or in the limit  $V \rightarrow \infty$

$$D(\mathbf{x}, x_0) = \frac{1}{(2\pi)^3} \int d^3 k \exp[i(\mathbf{k}\mathbf{x})] \frac{\sin k_0 x_0}{k_0}. \quad (11)$$

By to we understand the positive root

$$k_0 = +(k^2 + \kappa^2)^{1/2} \quad (12)$$

The  $D$  function is uniquely determined by the conditions:

$$\square D - \kappa^2 D = 0; \quad D(\mathbf{x}, 0) = 0;$$

$$\left( \frac{\partial D}{\partial x_0} \right)_{x_0=0} = \delta(\mathbf{x}). \quad (13)$$

For  $\kappa = 0$  we have simply

$$D(\mathbf{x}, x_0) = \{\delta(r - x_0) - \delta(r + x_0)\}/4\pi r. \quad (14)$$

This expression also determines the singularity of  $D(\mathbf{x}, x_0)$  on the light cone for  $\kappa \neq 0$ . But in the latter case  $D$  is no longer different from zero in the inner part of the cone. One finds for this region<sup>13</sup>

$$D(\mathbf{x}, x_0) = -\frac{1}{4\pi r} \frac{\partial}{\partial r} F(r, x_0)$$

---

<sup>12</sup>The consistent development of this method leads to the “many-time formalism” of Dirac, which has been given by P. A. M. Dirac, Quantum Mechanics (Oxford, second edition, 1935).

<sup>13</sup>See P. A. M. Dirac, Proc. Camb. Phil. Soc. **30**, 150 (1934).

with

$$F(r, x_0) = \begin{cases} J_0[\kappa(x_0^2 - r^2)^{1/2}] & \text{for } x_0 > r \\ 0 & \text{for } r > x_0 > -r \\ -J_0[\kappa(x_0^2 - r^2)^{1/2}] & \text{for } -r > x_0. \end{cases} \quad (15)$$

The jump from + to - of the function  $F$  on the light cone corresponds to the  $\delta$  singularity of  $D$  on this cone. For the following it will be of decisive importance that  $D$  vanish in the exterior of the cone (i.e., for  $r > x_0 > -r$ ).

The form of the factor  $d^3k/k_0$ , is determined by the fact that  $d^3k/k_0$  is invariant on the hyperboloid ( $k$ ) of the four-dimensional momentum space  $(\kappa, k_0)$ . It is for this reason that, apart from  $D$ , there exists just one more function which is invariant and which satisfies the wave equation (9), namely,

$$D_1(\mathbf{x}, x_0) = \frac{1}{(2\pi)^3} \int d^3k \exp[i(\mathbf{k}\mathbf{x})] \frac{\cos k_0 x_0}{k_0}. \quad (16)$$

For  $\kappa = 0$  one finds

$$D_1(\mathbf{x}, x_0) = \frac{1}{2\pi^2} \frac{1}{r^2 - x_0^2}. \quad (17)$$

In general it follows

$$D_1(\mathbf{x}, x_0) = \frac{1}{4\pi} \frac{1}{r} \frac{\partial}{\partial r} F_1(r, x_0)$$

$$F_1(r, x_0) = \begin{cases} N_0[\kappa(x_0^2 - r^2)^{1/2}] & \text{for } x_0 > r \\ -iH_0^{(1)}[i\kappa(r^2 - x_0^2)^{1/2}] & \text{for } r > x_0 > -r \\ N_0[\kappa(x_0^2 - r^2)^{1/2}] & \text{for } -r > x_0. \end{cases} \quad (18)$$

Here  $N_0$  stands for Neumann's function and  $H_0^{(1)}$  for the first Hankel cylinder function. The strongest singularity of  $D$ , on the surface of the light cone is in general determined by (17).

We shall, however, expressively postulate in the following *that all physical quantities at finite distances exterior to the light cone (for  $|x'_0 - x''_0| < |\mathbf{x}' - \mathbf{x}''|$ ) are commutable.*<sup>14</sup> It follows from this that the bracket expressions of all quantities which satisfy the force-free wave equation (9) can be expressed by the function  $D$  and (a finite number) of derivatives of it without using the function  $D_1$ . This is also true for brackets with the + sign,

---

<sup>14</sup>For the canonical quantization formalism this postulate is satisfied implicitly. But this postulate is much more general than the canonical formalism.

since otherwise it would follow that gauge invariant quantities, which are constructed bilinearly from the  $U^{(r)}$ , as for example the charge density, are noncommutable in two points with a space-like distance. <sup>15</sup>

The justification for our postulate lies in the fact that measurements at two space points with a space-like distance can never disturb each other, since no signals can be transmitted with velocities greater than that of light. Theories which would make use of the  $D_1$  function in their quantization would be very much different from the known theories in their consequences.

At once we are able to draw further conclusions about the number of derivatives of  $D$  function which can occur in the bracket expressions, if we take into account the invariance of the theories under the transformations of the restricted Lorentz group and if we use the results of the preceding section on the class division of the tensors. We assume the quantities  $U^{(r)}$  to be ordered in such a way that each field component is composed only of quantities of the same class. We consider especially the bracket expression of a field component  $U^{(r)}$  with its own complex conjugate

$$[U^{(r)}(\mathbf{x}', x'_0), U^{*(r)}(\mathbf{x}'', x''_0)].$$

We distinguish now the two cases of half-integral and integral spin. In the former case this expression transforms according to (8) under Lorentz transformations as a tensor of odd rank. In the second case, however, it transforms as a tensor of even rank. Hence we have for half-integral spin

$$\begin{aligned} & [U^{(r)}(\mathbf{x}', x'_0), U^{*(r)}(\mathbf{x}'', x''_0)] \\ &= \text{odd number of derivatives of the function} \\ & D(\mathbf{x}' - \mathbf{x}'', x'_0 - x''_0) \end{aligned} \tag{19a}$$

and similarly for integral spin

$$\begin{aligned} & [U^{(r)}(\mathbf{x}', x'_0), U^{*(r)}(\mathbf{x}'', x''_0)] \\ &= \text{even number of derivatives of the function} \\ & D(\mathbf{x}' - \mathbf{x}'', x'_0 - x''_0). \end{aligned} \tag{19b}$$

This must be understood in such a way that on the right-hand side there may occur a complicated sum of expressions of the type indicated. We consider now the following expression, which is symmetrical in the two points

$$X \equiv [U^{(r)}(\mathbf{x}', x'_0), U^{*(r)}(\mathbf{x}'', x''_0)] + [U^{(r)}(\mathbf{x}'', x''_0), U^{*(r)}(\mathbf{x}', x'_0)]. \tag{19}$$

---

<sup>15</sup>See W. Pauli, Ann. de l'Inst. H. Poincaré **6**, 137 (1936), esp. § 3.

Since the  $D$  function is even in the space coordinates odd in the time coordinate, which can be seen at once from Eqs. (11) or (15), it follows from the symmetry of  $X$  that  $X = \text{even number of space-like times odd numbers of time-like derivatives of } D(\mathbf{x}' - \mathbf{x}'', x'_0 - x''_0)$ . This is fully consistent with the postulate (19a) for half-integral spin, but in contradiction with (19b) for integral spin unless  $X$  vanishes. We have therefore the result for integral spin

$$[U^{(r)}(\mathbf{x}', x'_0), U^{*(r)}(\mathbf{x}'', x''_0)] + [U^{(r)}(\mathbf{x}'', x''_0), U^{*(r)}(\mathbf{x}', x'_0)] = 0. \quad (20)$$

So far we have not distinguished between the two cases of Bose statistics and the exclusion principle. In the former case, one has the ordinary bracket with the — sign, in the latter case, according to Jordan and Wigner, the bracket

$$[A, B]_+ = AB + BA$$

with the + sign. *By inserting the brackets with the + sign into (20) we have an algebraic contradiction*, since the left-hand side is essentially positive for  $x' = x''$  and cannot vanish unless both  $U^{(r)}$  and  $U^{*(r)}$  vanish.<sup>16</sup>

Hence we come to the result: *For integral spin the quantization according to the exclusion principle is not possible. For this result it is essential, that the use of the  $D_1$  function in place of the  $D$  function be, for general reasons, discarded.*

On the other hand, it is formally possible to quantize the theory for half-integral spins according to Einstein-Bose-statistics, *but according to the general result of the preceding section the energy of the system would not be positive*. Since for physical reasons it is necessary to postulate this, we must apply the exclusion principle in connection with Dirac's hole theory.

For the positive proof that a theory with a positive total energy is possible by quantization according to Bose-statistics (exclusion principle) for

---

<sup>16</sup>This contradiction may be seen also by resolving  $U^{(r)}$  into eigen vibrations according to

$$\begin{aligned} U^{*(r)}(\mathbf{x}, x_0) &= V^{-1/2} \sum_k \{U_+^*(k) \exp[i\{-(\mathbf{k}\mathbf{x}) + k_0 x_0\}] + U_-^*(k) \exp[i\{(\mathbf{k}\mathbf{x}) - k_0 x_0\}]\} \\ U^{(r)}(\mathbf{x}, x_0) &= V^{-1/2} \sum_k \{U_+(k) \exp[i\{(\mathbf{k}\mathbf{x}) - k_0 x_0\}] + U_-(k) \exp[i\{-(\mathbf{k}\mathbf{x}) + k_0 x_0\}]\}. \end{aligned}$$

The equation (21) leads then, among others, to the relation

$$[U_+^*(k), U_+(k)] + [U_-(k), U_-^*(k)] = 0,$$

a relation, which is not possible for brackets with the + sign unless  $U_{\pm}(k)$  and  $U_{\pm}^*(k)$  vanish.

integral (half-integral) spins, we must refer to the already mentioned paper by Fierz. In another paper by Fierz and Pauli <sup>17</sup> the case of an external electromagnetic field and also the connection between the special case of spin 2 and the gravitational theory of Einstein has been discussed. In conclusion we wish to state, that according to our opinion the connection between spin and statistics is one of the most important applications of the special relativity theory.

---

<sup>17</sup>M. Fierz and W. Pauli, Proc. Roy. Soc. **A173**, 211 (1939).

## Expanding Universe and the Origin of Elements

G. GAMOW

The George Washington University, Washington, D.C.

September 13, 1946

It is generally agreed at present that the relative abundances of various chemical elements were determined by physical conditions existing in the universe during the early stages of its expansion, when the temperature and density were sufficiently high to secure appreciable reaction-rates for the light as well as for the heavy nuclei.

In all the so-far published attempts in this direction the observed abundance-curve is supposed to represent some equilibrium state determined by nuclear binding energies at some very high temperature and density [1] [2] [3]. This point of view encounters, however, serious difficulties in the comparison with empirical facts. Indeed, since binding energy is, in a first approximation, a linear function of atomic weight, any such equilibrium theory would necessarily lead to a rapid exponential decrease of abundance through the entire natural sequence of elements. It is known, however, that whereas such a rapid decrease actually takes place for the first half of chemical elements, the abundance of heavier nuclei remains nearly constant [4]. Attempts have been made<sup>2</sup> to explain this discrepancy by the assumption that heavy elements were formed at higher temperatures, and that their abundances were already "frozen" when the adjustment of lighter elements was taking place. Such an explanation, however, can be easily ruled out if one remembers that at the temperatures and densities in question (about  $10^{10}$  K, and  $10^6$  g/cm<sup>3</sup>) nuclear transformations are mostly caused by the processes of absorption and re-evaporation of free neutrons so that their rates are essentially the same for the light and for the heavy elements. Thus it appears that the only way of explaining the observed abundance-curve

lies in the assumption of some kind of unequilibrium process taking place during a limited interval of time.

The above conclusion finds a strong support in the study of the expansion process itself. According to the general theory of expanding universe [5], the time dependence of any linear dimension  $l$  it is given by the formula

$$\frac{dl}{dt} = \left( \frac{8\pi G}{3} \rho l^2 - \frac{G^2}{R^2} \right)^{1/2} \quad (1)$$

where  $G$  is the Newton constant,  $\rho$  the mean density, and  $R$  (real or imaginary) a constant describing the curvature of space. It may be noticed that the above expression represents a relativistic analog of the familiar classical formula

$$v = \left( 2 \cdot \frac{4\pi l^2}{3} \rho \cdot \frac{G}{l} - 2E \right)^{1/2} \quad (2)$$

for the inertial expansion-velocity of a gravitating dust sphere with the total energy  $E$  per unit mass. The imaginary and real values of  $R$  correspond to an unlimited expansion (in case of superescape velocity), and to the expansion which will be ultimately turned into a contraction by the forces of gravity (subescape velocity). To use some definite numbers, let us consider in the present state of the universe (considered as quite uniform) a cube containing, say, 1 g of matter. Since the present mean density of the universe is  $\rho_{\text{present}} \approx 10^{-30}$  g/cm<sup>3</sup>, the side of our cube will be;  $l_{\text{present}} \approx 10^{10}$  cm. According to Hubble [6], the present expansion-rate of the universe is  $1.8 \times 10^{-17}$  cm/sec, per cm, so that  $(dl/dt)_{\text{present}} \approx 1.8 \times 10^{-7}$  cm/sec. Substituting the numerical values in (1) we obtain

$$1.8 \times 10^{-7} = (5.7 \times 10^{-17} - G^2/R^2)^{1/2}, \quad (3)$$

showing that at the present stage of expansion the first term under the radical (corresponding to the potential energy of gravity) is negligibly small as compared with the second one. For the numerical value of the (constant) radius of curvature we get from (3):  $R = 1.7 \times 10^{17} \sqrt{-1}$  cm or about 0.2 imaginary light year.

In the past history of the universe, when  $l$  was considerably smaller, and  $\rho$  correspondingly larger, the first term in (1) was playing an important role corresponding physically to the slowing-down effect of gravity on the original expansion. The transition from the slowed down to the free expansion took place at the epoch when the two terms were comparable, i.e., when  $l$  was about one thousandth of its present value. At this epoch the gravitational

clustering of matter into stars, stellar clusters, and galaxies, probably must have taken place [7].

Applying our formula (2) with  $G^2/R^2 = -3.3 \times 10^{-14}$  to the earlier epoch when the average density of masses in the universe was of the order of  $10^5$  g/cm<sup>3</sup> (as required by the conditions for the formation of elements), we find that at that time  $l \approx 10^{-2}$  cm, and  $dl/dt \approx 0.01$  cm/sec. This means that at *the epoch when the mean density of the universe was of the order of  $10^5$  g/cm<sup>3</sup>, the expansion must have been proceeding at such a high rate, that this high density was reduced by an order of magnitude in only about one second.* It goes without saying that one must be very careful in extrapolating the expansion formula to such an early epoch, but, on the other hand, this formula represents nothing more than the statement of the law of conservation of energy in the inertial expansion against the forces of gravity.

Returning to our problem of the formation of elements, we see that *the conditions necessary for rapid nuclear reactions were existing only for a very short time*, so that it may be quite dangerous to speak about an equilibrium-state which must have been established during this period. It is also interesting to notice that the calculated time-period during which rapid nuclear transformations could have taken place is considerably shorter than the  $\beta$ -decay period of free neutrons which is presumably of the order of magnitude of one hour. Thus if free neutrons were present in large quantities in the beginning of the expansion, the mean density and temperature of expanding matter must have dropped to comparatively low values *before* these neutrons had time to turn into protons. We can anticipate that neutrons forming this comparatively cold cloud were gradually coagulating into larger and larger neutral complexes which later turned into various atomic species by subsequent processes of  $\beta$ -emission. From this point of view the decrease of relative abundance along the natural sequence of elements must be understood as being caused by the longer time which was required for the formation of heavy neutronic complexes by the successive processes of radiative capture. The present high abundance of hydrogen must have resulted from the competition between the  $\beta$ -decay of original neutrons which was turning them into inactive protons, and the coagulation-process through which these neutrons were being incorporated into heavier nuclear units.

It is hoped that the further more detailed development of the ideas presented above will permit us to understand the observed abundance-curve of chemical elements giving at the same time valuable information concerning the early stages of the expanding universe.

## References

- [1] v. Weizsäcker, Physik. Zeits., **39**, 633 (1938).
- [2] Chandrasekhar and Henrich, Astrophys. J. **95**, 288 (1942).
- [3] G. Wataghin, Phys. Rev. **66**, 149 (1944).
- [4] Goldschmidt, Verleilung der Elemente (Oslo, 1938).
- [5] R. Tolman, Relativity, Thermodynamics and Cosmotology (Oxford Press, New York. 1934).
- [6] Hubble, The Realm of the Nebulos (Yale University Press. New Haven, 1936).
- [7] G. Gamow and E. Teller, Phys. Rev. **55**, 654 (1939).

## Space-Time Approach to Non-Relativistic Quantum Mechanics

R.P. Feynman  
Cornell University,  
Ithaca, New York

— — ◇ ◇ — —  
Reprinted in “Quantum Electrodynamics”, edited by Julian Schwinger  
— — ◇ ◇ — —

### Abstract

Non-relativistic quantum mechanics is formulated here in a different way. It is, however, mathematically equivalent to the familiar formulation. In quantum mechanics the probability of an event which can happen in several different ways is the absolute square of a sum of complex contributions, one from each alternative way. The probability that a particle will be found to have a path  $x(t)$  lying somewhere within a region of space time is the square of a sum of contributions, one from each path in the region. The contribution from a single path is postulated to be an exponential whose (imaginary) phase is the classical action (in units of  $\hbar$ ) for the path in question. The total contribution from all paths reaching  $x, t$  from the past is the wave function  $\psi(x, t)$ . This is shown to satisfy Schroedinger’s equation. The relation to matrix and operator algebra is discussed. Applications are indicated, in particular to eliminate the coordinates of the field oscillators from the equations of quantum electrodynamics.

### 1. Introduction

It is a curious historical fact that modern quantum mechanics began with two quite different mathematical formulations: the differential equation of

Schroedinger, and the matrix algebra of Heisenberg. The two, apparently dissimilar approaches, were proved to be mathematically equivalent. These two points of view were destined to complement one another and to be ultimately synthesized in Dirac's transformation theory.

This paper will describe what is essentially a third formulation of non-relativistic quantum theory. This formulation was suggested by some of Dirac's<sup>1</sup> <sup>2</sup> remarks concerning the relation of classical action<sup>3</sup> to quantum mechanics. A probability amplitude is associated with an entire motion of a particle as a function of time, rather than simply with a position of the particle at a particular time.

The formulation is mathematically equivalent to the more usual formulations. There are, therefore, no fundamentally new results. However, there is a pleasure in recognizing old things from a new point of view. Also, there are problems for which the new point of view offers a distinct advantage. For example, if two systems *A* and *B*, interact, the coordinates of one of the systems, say *B*, may be eliminated from the equations describing the motion of *A*. The interaction with *B* is represented by a change in the formula for the probability amplitude associated with a motion of *A*. It is analogous to the classical situation in which the effect of *B*, can be represented by a change in the equations of motion of *A* (by the introduction of terms representing forces acting on *A*). In this way the coordinates of the transverse, as well as of the longitudinal field oscillators, may be eliminated from the equations of quantum electrodynamics.

In addition, there is always the hope that the new point of view will inspire an idea for the modification of present theories, a modification necessary to encompass present experiments.

We first discuss the general concept of the superposition of probability amplitudes in quantum mechanics. We then show how this concept can be directly extended to define a probability amplitude for any motion or path (position *vs.* time) in space-time. The ordinary quantum mechanics is shown to result from the postulate that this probability amplitude has a phase proportional to the action, computed classically, for this path. This is true when the action is the time integral of a quadratic function of velocity. The relation to matrix and operator algebra is discussed in a way that

---

<sup>1</sup>P. A. M. Dirac, *The Principles of Quantum Mechanics* (The Clarendon Press, Oxford, 1935), second edition, Section 33; also, *Physik. Zeits. Sowjetunion* **3**, 64 (1933).

<sup>2</sup>P. A. M. Dirac, *Rev. Mod. Phys.* **17**, 195 (1945).

<sup>3</sup>Throughout this paper the term "action" will be used for the time integral of the Lagrangian along a path. When this path is the one actually taken by a particle, moving classically, the integral should more properly be called Hamilton's first principle function.

stays as close to the language of the new formulation as possible. There is no practical advantage to this, but the formulae are very suggestive if a generalization to a wider class of action functionals is contemplated. Finally, we discuss applications of the formulation. As a particular illustration, we show how the coordinates of a harmonic oscillator may be eliminated from the equations of motion of a system with which it interacts. This can be extended directly for application to quantum electrodynamics. A formal extension which includes the effects of spin and relativity is described.

## 2. The Superposition of Probability Amplitudes

The formulation to be presented contains as its essential idea the concept of a probability amplitude associated with a completely specified motion as a function of time. It is, therefore, worthwhile to review in detail the quantum-mechanical concept of the superposition of probability amplitudes. We shall examine the essential changes in physical outlook required by the transition from classical to quantum physics.

For this purpose, consider an imaginary experiment in which we can make three measurements successive in time: first of a quantity  $A$ , then of  $B$ , and then of  $C$ . There is really no need for these to be of different quantities, and it will do just as well if the example of three successive position measurements is kept in mind. Suppose that  $a$  is one of a number of possible results which could come from measurement  $A$ ,  $b$  is a result that could arise from  $B$ , and  $c$  is a result possible from the third measurement  $C$ .<sup>4</sup> We shall assume that the measurements  $A$ ,  $B$ , and  $C$  are the type of measurements that completely specify a state in the quantum-mechanical case. That is, for example, the state for which  $B$  has the value  $b$  is not degenerate.

It is well known that quantum mechanics deals with probabilities, but naturally this is not the whole picture. In order to exhibit, even more clearly, the relationship between classical and quantum theory, we could suppose that classically we are also dealing with probabilities but that all probabilities either are zero or one. A better alternative is to imagine in the classical case that the probabilities are in the sense of classical statistical mechanics (where, possibly, internal coordinates are not completely specified).

We define  $P_{ab}$  as the probability that if measurement  $A$  gave the result  $a$ ,

---

<sup>4</sup>For our discussion it is not important that certain values of  $a$ ,  $b$ , or  $c$  might be excluded by quantum mechanics but not by classical mechanics. For simplicity, assume the values are the same for both but that the probability of certain values may be zero.

then measurement  $B$  will give the result  $b$ . Similarly,  $P_{bc}$  is the probability that if measurement  $B$  gives the result  $b$ , then measurement  $C$  gives  $c$ . Further, let  $P_{ac}$  be the chance that if  $A$  gives  $a$ , then  $C$  gives  $c$ . Finally, denote by  $P_{abc}$  the probability of all three, i.e., if  $A$  gives  $a$ , then  $B$  gives  $b$ , and  $C$  gives  $c$ . If the events between  $a$  and  $b$  are independent of those between  $b$  and  $c$ , then

$$P_{abc} = P_{ab}P_{bc}. \quad (1)$$

This is true according to quantum mechanics when the statement that  $B$  is  $b$  is a complete specification of the state.

In any event, we expect the relation

$$P_{ac} = \sum_b P_{abc}. \quad (2)$$

This is because, if initially measurement  $A$  gives  $a$  and the system is later found to give the result  $c$  to measurement  $C$  quantity  $B$  must have had some value at the time intermediate to  $A$  and  $C$ . The probability that it was  $b$  is  $P_{abc}$ . We sum, or integrate, over all the mutually exclusive alternatives for  $b$  (symbolized by  $\sum_b$ ).

Now, the essential difference between classical and quantum physics lies in Eq. (2). In classical mechanics it is always true. In quantum mechanics it is often false. We shall denote the quantum-mechanical probability that a measurement of  $C$  results in  $c$  when it follows a measurement of  $A$  giving  $a$  by  $P_{ac}^q$ . Equation (2) is replaced in quantum mechanics by this remarkable law:<sup>5</sup> There exist complex numbers  $\varphi_{ab}, \varphi_{bc}, \varphi_{ac}$  such that

$$P_{ab} = |\varphi_{ab}|^2, \quad P_{bc} = |\varphi_{bc}|^2, \quad \text{and} \quad P_{ac}^q = |\varphi_{ac}|^2. \quad (3)$$

The classical law, obtained by combining (1) and (2),

$$P_{ac} = \sum_b P_{ab}P_{bc} \quad (4)$$

is replaced by

$$\varphi_{ac} = \sum_b \varphi_{ab}\varphi_{bc}. \quad (5)$$

If (5) is correct, ordinarily (4) is incorrect. The logical error made in deducing (4) consisted, of course, in assuming that to get from  $a$  to  $c$  the system

---

<sup>5</sup>We have assumed  $b$  is a non-degenerate state, and that therefore (1) is true. Presumably, if in some generalization of quantum mechanics (1) were not true, even for pure states  $b$ , (2) could be expected to be replaced by: There are complex numbers  $\varphi_{abc}$  such that  $P_{abc} = |\varphi_{abc}|^2$ . The analog of (5) is then  $\varphi_{ac} = \sum_b \varphi_{abc}$

had to go through a condition such that  $B$  had to have some definite value,  $b$ .

If an attempt is made to verify this, i.e., if  $B$  is measured between the experiments  $A$  and  $C$ , then formula (4) is, in fact, correct. More precisely, if the apparatus to measure  $B$  is set up and used, but no attempt is made to utilize the results of the  $B$  measurement in the sense that only the  $A$  to  $C$  correlation is recorded and studied, then (4) is correct. This is because the  $B$  measuring machine has done its job; if we wish, we could read the meters at any time without disturbing the situation any further. The experiments which gave  $a$  and  $c$  can, therefore, be separated into groups depending on the value of  $b$ .

Looking at probability from a frequency point of view (4) simply results from the statement that in each experiment giving  $a$  and  $c$ ,  $B$  had some value. The only way (4) could be wrong is the statement, " $B$  had some value," must sometimes be meaningless. Noting that (5) replaces (4) only under the circumstance that we make no attempt to measure  $B$ , we are led to say that the statement, " $B$  had some value," may be meaningless whenever we make no attempt to measure  $B$ <sup>6</sup>.

Hence, we have different results for the correlation of  $a$  and  $c$ , namely, Eq. (4) or Eq. (5), depending upon whether we do or do not attempt to measure  $B$ . No matter how subtly one tries, the attempt to measure  $B$  must disturb the system, at least enough to change the results from those given by (5) to those of (4)<sup>7</sup>. That measurements do, in fact, cause the necessary disturbances, and that, essentially, (4) could be false was first clearly enunciated by Heisenberg in his uncertainty principle. The law (5) is a result of the work of Schroedinger, the statistical interpretation of Born and Jordan, and the transformation theory of Dirac.<sup>8</sup>

Equation (5) is a typical representation of the wave nature of matter.

---

<sup>6</sup>It does not help to point out that we could have measured  $B$  had we wished. The fact is that we did not.

<sup>7</sup>How (4) actually results from (5) when measurements disturb the system has been studied particularly by J. von Neumann (*Mathematische Grundlagen der Quantenmechanik* (Dover Publications, New York, 1943)). The effect of perturbation of the measuring equipment is effectively to change the phase of the interfering components, by  $\theta_b$ , say, so that (5) becomes  $\varphi_{ac} = \sum_b e^{i\theta_b} \varphi_{ab} \varphi_{bc}$ . However, as von Neumann shows, the phase shifts must remain unknown if  $B$  is measured so that the resulting probability  $P_{ac}$  is the square of  $\varphi_{ac}$  averaged over all phases,  $\theta_b$ . This results in (4).

<sup>8</sup>If  $\mathbf{A}$  and  $\mathbf{B}$  are the operators corresponding to measurements  $A$  and  $B$ , and if  $\psi_a$ , and  $\psi_b$  are solutions of  $\mathbf{A}\psi_a = a\psi_a$ , and  $\mathbf{B}\chi_b = b\chi_b$ , then  $\varphi_{ab} = \int \chi_b^* \psi_a dx = (\chi_b^*, \psi_a)$ . Thus,  $\psi_{ab}$  is an element  $(a|b)$  of the transformation matrix for the transformation from a representation in which  $\mathbf{A}$  is diagonal to one in which  $\mathbf{B}$  is diagonal.

Here, the chance of finding a particle going from  $a$  to  $c$  through several different routes (values of  $b$ ) may, if no attempt is made to determine the route, be represented as the square of a sum of several complex quantities—one for each available route.

Probability can show the typical phenomena of interference, usually associated with waves, whose intensity is given by the square of the sum of contributions from different sources. The electron acts as a wave, (5), so to speak, as long as no attempt is made to verify that it is a particle; yet one can determine, if one wishes, by what route it travels just as though it were a particle; but when one does that, (4) applies and it does act like a particle.

These things are, of course, well known. They have already been explained many times.<sup>9</sup> However, it seems worth while to emphasize the fact that they are all simply direct consequences of Eq. (5), for it is essentially Eq. (5) that is fundamental in my formulation of quantum mechanics.

The generalization of Eqs. (4) and (5) to a large number of measurements, say  $A, B, C, D, \dots, K$ , is, of course, that the probability of the sequence  $a, b, c, d, \dots, k$ , is

$$P_{abcd\dots k} = |\varphi_{abcd\dots k}|^2.$$

The probability of the result  $a, c, k$ , for example, if  $b, d, \dots$  are measured, is the classical formula:

$$P_{ack} = \sum_b \sum_d \dots P_{abcd\dots k}, \quad (6)$$

while the probability of the same sequence  $a, c, k$  if no measurements are made between  $A$  and  $C$  and between  $C$  and  $K$  is

$$P_{ack}^q = \left| \sum_b \sum_d \dots \varphi_{abcd\dots k} \right|^2. \quad (7)$$

The quantity  $\varphi_{abcd\dots k}$  we can call the probability amplitude for the condition  $A = a, B = b, C = c, D = d, \dots, K = k$ . (It is, of course, expressible as a product  $\varphi_{ab}\varphi_{bc}\varphi_{cd} \dots \varphi_{jk}$ .)

### 3. The Probability Amplitude for a Space-Time Path

The physical ideas of the last section may be readily extended to define a probability amplitude for a particular completely specified space-time

---

<sup>9</sup>See, for example, W. Heisenberg, *The Physical Principles of the Quantum Theory* (University of Chicago Press, Chicago, 1930). particularly Chapter IV.

path. To explain how this may be done, we shall limit ourselves to a one-dimensional problem, as the generalization to several dimensions is obvious.

Assume that we have a particle which can take up various values of a coordinate  $x$ . Imagine that we make an enormous number of successive position measurements, let us say separated by a small time interval  $\epsilon$ . Then a succession of measurements such as  $A, B, C, \dots$  might be the succession of measurements of the coordinate  $x$  at successive times  $t_1, t_2, t_3, \dots$ , where  $t_{i+1} = t_i + \epsilon$ . Let the value, which might result from measurement of the coordinate at time  $t_i$  be  $x_i$ . Thus, if  $A$  is a measurement of  $x$  at  $t_1$  then  $x_1$  is what we previously denoted by  $a$ . From a classical point of view, the successive values,  $x_1, x_2, x_3, \dots$  of the coordinate practically define a path  $x(t)$ . Eventually, we expect to go the limit  $\epsilon \rightarrow 0$ .

The probability of such a path is a function of  $x_1, x_2, \dots, x_i, \dots$ , say  $P(\dots x_i, x_{i+1}, \dots)$ . The probability that the path lies in a particular region  $R$  of space-time is obtained classically by integrating  $P$  over that region. Thus, the probability that  $x_i$ , lies between  $a_i$  and  $b_i$  and  $x_{i+1}$  lies between  $a_{i+1}$  and etc., is

$$\begin{aligned} & \dots \int_{a_i}^{b_i} \int_{a_{i+1}}^{b_{i+1}} \dots P(\dots x_i, x_{i+1}, \dots) \dots dx, dx_{i+1} \dots = \\ & = \int_R P(\dots x_i, x_{i+1}, \dots) \dots dx, dx_{i+1} \dots, \end{aligned} \quad (8)$$

the symbol  $\int_R$  meaning that the integration is to be taken over those ranges of the variables which lie within the region  $R$ . This is simply Eq. (6) with  $a, b, \dots$  replaced by  $x_1, x_2, \dots$  and integration replacing summation.

In quantum mechanics this is the correct formula for the case that  $x_1, x_2, \dots, x_i, \dots$  were actually all measured, and then only those paths lying within  $R$  were taken. We would expect the result to be different if no such detailed measurements had been performed. Suppose a measurement is made which is capable only of determining that the path lies somewhere within  $R$ .

The measurement is to be what we might call an “ideal measurement.” We suppose that no further details could be obtained from the same measurement without further disturbance to the system. I have not been able to find a precise definition. We are trying to avoid the extra uncertainties that must be averaged over if, for example, more information were measured but not utilized. We wish to use Eq. (5) or (7) for all  $x_i$  and have no residual part to sum over in the manner of Eq. (4).

We expect that the probability that the particle is found by our “ideal measurement” to be, indeed, in the region  $R$  is the square of a complex number  $|\varphi(R)|^2$ . The number  $\varphi(R)$ , which we may call the probability amplitude for region  $R$  is given by Eq. (7) with  $a, b, \dots$  replaced by  $x_i, x_{i+1}, \dots$  and summation replaced by integration:

$$\varphi(R) = \lim_{\epsilon \rightarrow 0} \int_R \times \Phi(\dots x_i, x_{i+1} \dots) \dots dx_i dx_{i+1} \dots \quad (9)$$

The complex number  $\Phi(\dots x_i, x_{i+1} \dots)$  is a function of the variables  $x_i$ , defining the path. Actually, we imagine that the time spacing  $\epsilon$  approaches zero so that  $\Phi$  essentially depends on the entire path  $x(t)$  rather than only on just the values of  $x_i$ , at the particular times  $t_i$ ,  $x_i = x(t_i)$ . We might call  $\Phi$  the probability amplitude functional of paths  $x(t)$ .

We may summarize these ideas in our first postulate:

*I. If an ideal measurement is performed, to determine whether a particle has a path lying in a region of space-time, then the probability that the result will be affirmative is the absolute square of a sum of complex contributions, one from each path in the region.*

The statement of the postulate is incomplete. The meaning of a sum of terms one for “each” path is ambiguous. The precise meaning given in Eq. (9) is this: A path is first defined only by the positions  $x_i$ ; through which it goes at a sequence of equally spaced times, <sup>10</sup>  $t_i = t_{i-1} + \epsilon$ . Then all values of the coordinates within  $R$  have an equal weight. The actual magnitude of the weight depends upon  $\epsilon$  and can be so chosen that the probability of an event which is certain shall be normalized to unity. It may not be best to do so, but we have left this weight factor in a proportionality constant in the second postulate. The limit  $\epsilon \rightarrow 0$  must be taken at the end of a calculation.

When the system has several degrees of freedom the coordinate space  $x$  has several dimensions so that the symbol  $x$  will represent a set of coordinates  $(x^{(1)}, x^{(2)}, \dots, x^{(k)})$  for a system with  $k$  degrees of freedom. A path is a sequence of configurations for successive times and is described by giving the configuration  $x_i$ , or  $(x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(k)})$ , i.e., the value of each of the  $k$  coordinates for each time  $t_i$ . The symbol  $dx_i$ , will be understood to mean

---

<sup>10</sup>There are very interesting mathematical problems involved in the attempt to avoid the subdivision and limiting processes. Some sort of complex measure is being associated with the space of functions  $x(t)$ . Finite results can be obtained under unexpected circumstances because the measure is not positive everywhere, but the contributions from most of the paths largely cancel out. These curious mathematical problems are sidestepped by the subdivision process. However, one feels as Cavalieri must have felt calculating the volume of a pyramid before the invention of calculus.

the volume element in  $k$  dimensional configuration space (at time  $t_i$ ). The statement of the postulates is independent of the coordinate system which is used.

The postulate is limited to defining the results of position measurements. It does not say what must be done to define the result of a momentum measurement, for example. This is not a real limitation, however, because in principle the measurement of momentum of one particle can be performed in terms of position measurements of other particles, e.g., meter indicators. Thus, an analysis of such an experiment will determine what it is about the first particle which determines its momentum.

#### 4. The Calculation of the Probability Amplitude for a Path

The first postulate prescribes the type of mathematical framework required by quantum mechanics for the calculation of probabilities. The second postulate gives a particular content to this framework by prescribing how to compute the important quantity  $\Phi$  for each path:

*II. The paths contribute equally in magnitude, but the phase of their contribution is the classical action (in units of  $\hbar$ ); i.e., the time integral of the Lagrangian taken along the path.*

That is to say, the contribution  $\Phi[x(t)]$  from a given path  $x(t)$  is proportional to  $\exp(i/\hbar S[x(t)])$ , where the action  $S[x(t)] = \int L(\dot{x}(t), x(t))dt$  is the time integral of the classical Lagrangian  $L(\dot{x}, x)$  taken along the path in question. The Lagrangian, which may be an explicit function of the time, is a function of position and velocity. If we suppose it to be a quadratic function of the velocities, we can show the mathematical equivalence of the postulates here and the more usual formulation of quantum mechanics.

To interpret the first postulate it was necessary to define a path by giving only the succession of points  $x_i$ , through which the path passes at successive times  $t_i$ . To compute  $S = \int L(\dot{x}, x)dt$  we need to know the path at all points, not just at  $x_i$ . We shall assume that the function  $x(t)$  in the interval between  $t_i$  and  $t_{i+1}$  is the path followed by a classical particle, with the Lagrangian  $L$ , which starting from  $x_i$ , at  $t_i$  reaches  $x_{i+1}$  at  $t_{i+1}$ . This assumption is required to interpret the second postulate for discontinuous paths. The quantity  $\Phi(\dots x_i, x_{i+1}, \dots)$  can be normalized (for various  $\epsilon$ ) if desired, so that the probability of an event which is certain is normalized to unity as  $\epsilon \rightarrow 0$ .

There is no difficulty in carrying out the action integral because of the

sudden changes of velocity encountered at the times  $t_i$  as long as  $L$  does not depend upon any higher time derivatives of the position than the first. Furthermore, unless  $L$  is restricted in this way the end points are not sufficient to define the classical path. Since the classical path is the one which makes the action a minimum, we can write

$$S = \sum_i S(x_{i+1}, x_i), \quad (10)$$

where

$$S(x_{i+1}, x_i) = \text{Min.} \int_{t_i}^{t_{i+1}} L(\dot{x}(t), x(t)) dt. \quad (11)$$

Written in this way, the only appeal to classical mechanics is to supply us with a Lagrangian function. Indeed, one could consider postulate two as simply saying, “ $\Phi$  is the exponential of  $i$  times the integral of a real function of  $x(t)$  and its first time derivative.” Then the classical equations of motion might be derived later as the limit for large dimensions. The function of  $x$  and  $\dot{x}$  then could be shown to be the classical Lagrangian within a constant factor.

Actually, the sum in (10), even for finite  $\epsilon$  is infinite and hence meaningless (because of the infinite extent of time). This reflects a further incompleteness of the postulates. We shall have to restrict ourselves to a finite, but arbitrarily long, time interval.

Combining the two postulates and using Eq. (10). we find

$$\varphi(R) = \lim_{\epsilon \rightarrow 0} \int_R \times \exp \left[ \frac{i}{\hbar} \sum_i S(x_{i+1}, x_i) \right] \cdots \frac{dx_{i+1}}{A} \frac{dx_i}{A} \cdots, \quad (12)$$

where we have let the normalization factor be split into a factor  $1/A$  (whose exact value we shall presently determine) for each instant of time. The integration is just over those values  $x_i, x_{i+1}, \dots$  which lie in the region  $R$ . This equation, the definition (11) of  $S(x_{i+1}, x_i)$ , and the physical interpretation of  $|\varphi(R)|^2$  as the probability that the particle will be found in  $R$ , complete our formulation of quantum mechanics.

## 5. Definition of the Wave Function

We now proceed to show the equivalence of these postulates to the ordinary formulation of quantum mechanics. This we do in two steps. We show in this

section how the wave function may be defined from the new point of view. In the next section we shall show that this function satisfies Schroedinger's differential wave equation.

We shall see that it is the possibility, (10), of expressing  $S$  as a sum, and hence  $\Phi$  as a product, of contributions from successive sections of the path, which leads to the possibility of defining a quantity having the properties of a wave function.

To make this clear, let us imagine that we choose a particular time  $t$  and divide the region  $R$  in Eq. (12) into pieces, future and past relative to  $t$ . We imagine that  $R$  can be split into: (a) a region  $R'$ , restricted in any way in space, but lying entirely earlier in time than some  $t'$ , such that  $t' < t$ ; (b) a region  $R''$  arbitrarily restricted in space but lying entirely later in time than  $t''$ , such that  $t'' > t$ ; (c) the region between  $t'$  and  $t''$  in which all the values of  $x$  coordinates are unrestricted, i.e., all of space-time between  $t'$  and  $t''$ . The region (c) is not absolutely necessary. It can be taken as narrow in time as desired. However, it is convenient in letting us consider varying  $t$  a little without having to redefine  $R'$  and  $R''$ . Then  $|\varphi(R', R'')|^2$  is the probability that the path occupies  $R'$  and  $R''$ . Because  $R'$  is entirely previous to  $R''$ , considering the time  $t$  as the present, we can express this as the probability that the path had been in region  $R'$  and will be in region  $R''$ . If we divide by a factor, the probability that the path is in  $R'$ , to renormalize the probability we find:  $|\varphi(R', R'')|^2$  is the (relative) probability that if the system were in region  $R'$  it will be found later in  $R''$ .

This is, of course, the important quantity in predicting the results of many experiments. We prepare the system in a certain way (e.g., it was in region  $R'$ ) and then measure some other property (e.g., will it be found in region  $R''$ ?). What does (12) say about computing this quantity, or rather the quantity  $\varphi(R', R'')$  of which it is the square?

Let us suppose in Eq. (12) that the time  $t$  corresponds to one particular point  $k$  of the subdivision of time into steps  $\epsilon$ , i.e., assume  $t = t_k$ , the index  $k$ , of course, depending upon the subdivision  $\epsilon$ . Then, the exponential being the exponential of a sum may be split into a product of two factors

$$\exp \left[ \frac{i}{\hbar} \sum_{i=k}^{\infty} S(x_{i+1}, x_i) \right] \cdot \exp \left[ \frac{i}{\hbar} \sum_{i=-\infty}^{k-1} S(x_{i+1}, x_i) \right]. \quad (13)$$

The first factor contains only coordinates with index  $k$  or higher, while the second contains only coordinates with index  $k$  or lower. This split is possible because of Eq. (10), which results essentially from the fact that the Lagrangian is a function only of positions and velocities. First, the

integration on all variables  $x_i$  for  $i > k$  can be performed on the first factor resulting in a function of  $x_k$  (times the second factor). Next, the integration on all variables  $x_i$ , for  $i < k$  can be performed on the second factor also, giving a function of  $x_k$ . Finally, the integration on  $x_k$  can be performed. That is,  $\varphi(R', R'')$  can be written as the integral over  $x_k$  of the product of two factors. We will call these  $\chi^*(x_k, t)$  and  $\psi(x_k, t)$ :

$$\varphi(R', R'') = \int \chi^*(x, t)\psi(x, t)dx, \quad (14)$$

where

$$\psi(x_k, t) = \text{Lim}_{\epsilon \rightarrow 0} \int_{R'} \times \exp \left[ \frac{i}{\hbar} \sum_{i=-\infty}^{k-1} S(x_{i+1}, x_i) \right] \frac{dx_{k-1}}{A} \frac{dx_{k-2}}{A} \dots, \quad (15)$$

and

$$\chi^*(x_k, t) = \text{Lim}_{\epsilon \rightarrow 0} \int_{R''} \exp \left[ \frac{i}{\hbar} \sum_{i=k}^{\infty} S(x_{i+1}, x_i) \right] \cdot \frac{1}{A} \frac{dx_{k+1}}{A} \frac{dx_{k+2}}{A} \dots \quad (16)$$

The symbol  $R'$  is placed on the integral for  $\psi$  to indicate that the coordinates are integrated over the region  $R'$ , and, for  $t_i$  between  $t'$  and  $t$ , over all space. In like manner, the integral for  $\chi^*$  is over  $R''$  and over all space for those coordinates corresponding to times between  $t$  and  $t''$ . The asterisk on  $\chi^*$  denotes complex conjugate, as it will be found more convenient to define (16) as the complex conjugate of some quantity,  $\chi$ .

The quantity  $\psi$  depends only upon the region  $R'$  previous to  $t$ , and is completely denned if that region is known. It does not depend, in any way, upon what will be done to the system after time  $t$ . This latter information is contained in  $\chi$ . Thus, with  $\psi$  and  $\chi$  we have separated the past history from the future experiences of the system. This permits us to speak of the relation of past and future in the conventional manner. Thus, if a particle has been in a region of space-time  $R'$  it may at time  $t$  be said to be in a certain condition, or state, determined only by its past and described by the so-called wave function  $\psi(x, t)$ . This function contains all that is needed to predict future probabilities. For, suppose, in another situation, the region  $R'$  were different, say  $r'$ , and possibly the Lagrangian for times before  $t$  were also altered. But, nevertheless, suppose the quantity from Eq. (15) turned out to be the same. Then, according to (14) the probability of ending in any region  $R''$  is the same for  $R'$  as for  $r'$ . Therefore, future measurements will not distinguish whether the system had occupied  $R'$  or  $r'$ . Thus, the

wave function  $\psi(x, t)$  is sufficient to define those attributes which are left from past history which determine future behavior.

Likewise, the function  $\chi(x, t)$  characterizes the experience, or, let us say, experiment to which the system is to be subjected. If a different region,  $r''$  and different Lagrangian after  $t$ , were to give the same  $\chi^*(x, t)$  via Eq. (16), as does region  $R''$ , then no matter what the preparation,  $\psi$ , Eq. (14) says that the chance of finding the system in  $R''$  is always the same as finding it in  $r''$ . The two “experiments”  $R''$  and  $r''$  are equivalent, as they yield the same results. We shall say loosely that these experiments are to determine with what probability the system is in state  $\chi$ . Actually, this terminology is poor. The system is really in state  $\psi$ . The reason we can associate a state with an experiment is, of course, that for an ideal experiment there turns out to be a unique state (whose wave function is  $\chi(x, t)$ ). for which the experiment succeeds with certainty.

Thus, we can say: the probability that a system in state  $\psi$  will be found by an experiment whose characteristic state is  $\chi$  (or, more loosely, the chance that a system in state  $\psi$  will appear to be in  $\chi$ ) is

$$\left| \int \chi^*(x, t) \psi(x, t) dx \right|^2. \quad (17)$$

These results agree, of course, with the principles of ordinary quantum mechanics. They are a consequence of the fact that the Lagrangian is a function of position, velocity, and time only.

## 6. The Wave Equation

To complete the proof of the equivalence with the ordinary formulation we shall have to show that the wave function defined in the previous section by Eq. (15) actually satisfies the Schroedinger wave equation. Actually, we shall only succeed in doing this when the Lagrangian  $L$  in (11) is a quadratic, but perhaps inhomogeneous, form in the velocities  $\dot{x}(t)$ . This is not a limitation, however, as it includes all the cases for which the Schroedinger equation has been verified by experiment.

The wave equation describes the development of the wave function with time. We may expect to approach it by noting that, for finite  $\epsilon$ , Eq. (15) permits a simple recursive relation to be developed. Consider the appearance of Eq. (15) if we were to compute  $\psi$  at the next instant of time:

$$\psi(x_{k+1}, t + \epsilon) = \int_{R'} \exp \left[ \frac{i}{\hbar} \sum_{i=-\infty}^k S(x_{i+1}, x_i) \right] \times \frac{dx_k}{A} \frac{dx_{k-1}}{A} \dots \quad (15')$$

This is similar to (15) except for the integration over the additional variable  $x_k$ , and the extra term in the sum in the exponent. This term means that the integral of (15') is the same as the integral of (15) except for the factor  $(1/A)\exp(i/\hbar)S(x_{k+1}, x_k)$ . Since this does not contain any of the variables  $x_i$ , for  $i$  less than  $k$ , all of the integrations on  $dx$ , up to  $dx_{k-1}$  can be performed with this factor left out. However, the result of these integrations is by (15) simply  $\psi(x_k, t)$ . Hence, we find from (15') the relation

$$\psi(x_{k+1}, t + \epsilon) = \int \exp\left[\frac{i}{\hbar}S(x_{k+1}, x_k)\right] \psi(x_k, t) dx_k / A. \quad (18)$$

This relation giving the development of  $\psi$  with time will be shown, for simple examples, with suitable choice of  $A$ , to be equivalent to Schroedinger's equation. Actually, Eq. (18) is not exact, but is only true in the limit  $\epsilon \rightarrow 0$  and we shall derive the Schroedinger equation by assuming (18) is valid to first order in  $\epsilon$ . The Eq. (18) need only be true for small  $\epsilon$  to the first order in  $\epsilon$ . For if we consider the factors in (15) which carry us over a finite interval of time,  $T$ , the number of factors is  $T/\epsilon$ . If an error of order  $\epsilon^2$  is made in each, the resulting error will not accumulate beyond the order  $\epsilon^2(T/\epsilon)$  or  $T\epsilon$ , which vanishes in the limit.

We shall illustrate the relation of (18) to Schroedinger's equation by applying it to the simple case of a particle moving in one dimension in a potential  $V(x)$ . Before we do this, however, we would like to discuss some approximations to the value  $S(x_{i+1}, x_i)$  given in (11) which will be sufficient for expression (18).

The expression defined in (11) for  $S(x_{i+1}, x_i)$  is difficult to calculate exactly for arbitrary  $\epsilon$  from classical mechanics. Actually, it is only necessary that an approximate expression for  $S(x_{i+1}, x_i)$  be used in (18), provided the error of the approximation be of an order smaller than the first in  $\epsilon$ . We limit ourselves to the case that the Lagrangian is a quadratic, but perhaps inhomogeneous, form in the velocities  $\dot{x}(t)$ . As we shall see later, the paths which are important are those for which  $x_{i+1} - x_i$  is of order  $\epsilon^{1/2}$ . Under these circumstances, it is sufficient to calculate the integral in (11) over the classical path taken by a free particle.<sup>11</sup> In *Cartesian coordinates*<sup>12</sup> the path of a free particle is a straight line so the integral of (11) can be taken

---

<sup>11</sup>It is assumed that the "forces" enter through a scalar and vector potential and not in terms involving the square of the velocity. More generally, what is meant by a free particle is one for which the Lagrangian is altered by omission of the terms linear in, and those independent of, the velocities.

<sup>12</sup>More generally, coordinates for which the terms quadratic in the velocity in  $L(\dot{x}, x)$  appear with constant coefficients.

along a straight line. Under these circumstances it is sufficiently accurate to replace the integral by the trapezoidal rule

$$S(x_{i+1}, x_i) = \frac{\epsilon}{2} L\left(\frac{x_{i+1} - x_i}{\epsilon}, x_{i+1}\right) + \frac{\epsilon}{2} L\left(\frac{x_{i+1} - x_i}{\epsilon}, x_i\right) \quad (19)$$

or, if it proves more convenient,

$$S(x_{i+1}, x_i) = \epsilon L\left(\frac{x_{i+1} - x_i}{\epsilon}, \frac{x_{i+1} + x_i}{2}\right). \quad (20)$$

These are not valid in a general coordinate system, e.g., spherical. An even simpler approximation may be used if, in addition, there is no vector potential or other terms linear in the velocity (see page 376):

$$S(x_{i+1}, x_i) = \epsilon L\left(\frac{x_{i-1} - x_i}{\epsilon}, x_{i+1}\right). \quad (21)$$

Thus, for the simple example of a particle of mass  $m$  moving in one dimension under a potential  $V(x)$ , we can set

$$S(x_{i+1}, x_i) = \frac{m\epsilon}{2} \left( \frac{x_{i+1} - x_i}{\epsilon} \right)^2 - \epsilon V(x_{i+1}). \quad (22)$$

For this example, then, Eq. (18) becomes

$$\begin{aligned} \psi(x_{k+1}, t + \epsilon) = \int \exp & \left[ \frac{i\epsilon}{\hbar} \left\{ \frac{m}{2} \left( \frac{x_{k+1} - x_k}{\epsilon} \right)^2 - \right. \right. \\ & \left. \left. - V(x_{k+1}) \right] \psi(x_k, t) dx_k / A. \right. \end{aligned} \quad (23)$$

Let us call  $x_{k+1} = x$  and  $x_{k+1} - x_k = \xi$  so that  $x_k = x - \xi$ . Then (23) becomes

$$\psi(x, t + \epsilon) = \int \exp \frac{im\xi^2}{\epsilon \cdot 2\hbar} \cdot \exp \frac{-i\epsilon V(x)}{\hbar} \cdot \psi(x - \xi, t) \frac{d\xi}{A}. \quad (24)$$

The integral on  $\xi$  will converge if  $\psi(x, t)$  falls off sufficiently for large  $x$  (certainly if  $\int \psi^*(x)\psi(x)dx = 1$ ). In the integration on  $\xi$ , since  $\epsilon$  is very small, the exponential of  $im\xi^2/2\hbar\epsilon$  oscillates extremely rapidly except in the region about  $\xi = 0$  ( $\xi$  of order  $(\hbar\epsilon/m)^{1/2}$ ). Since the function  $\psi(x - \xi, t)$  is a relatively smooth function of  $\xi$  (since  $\epsilon$  may be taken as small as desired), the region where the exponential oscillates rapidly will contribute very little

because of the almost complete cancelation of positive and negative contributions. Since only small  $\xi$  are effective,  $\psi(x - \xi, t)$  may be expanded as a Taylor series. Hence,

$$\begin{aligned} \psi(x, t + \epsilon) &= \exp\left(\frac{-i\epsilon V(x)}{\hbar}\right) \times \\ &\times \int \exp\left(\frac{im\xi^2}{2\hbar\epsilon}\right) \left[ \psi(x, t) - \xi \frac{\partial\psi(x, t)}{\partial x} + \frac{\xi^2}{2} \frac{\partial^2\psi(x, t)}{\partial x^2} - \dots \right] d\xi / A. \end{aligned} \quad (25)$$

Now

$$\begin{aligned} \int_{-\infty}^{\infty} \exp(im\xi^2/2\hbar\epsilon) d\xi &= (2\pi\hbar\epsilon i/m)^{1/2}, \\ \int_{-\infty}^{\infty} \exp(im\xi^2/2\hbar\epsilon) \xi d\xi &= 0, \\ \int_{-\infty}^{\infty} \exp(im\xi^2/2\hbar\epsilon) \xi^2 d\xi &= (\hbar\epsilon i/m)(2\pi\hbar\epsilon i/m)^{1/2}, \end{aligned} \quad (26)$$

while the integral containing  $\xi^2$  is zero, for like the one with  $\xi$  it possesses an odd integrand, and the ones with  $\xi^4$  are of at least the order  $\epsilon$  smaller than the ones kept here.<sup>13</sup> If we expand the left-hand side to first order in  $\epsilon$  (25) becomes

$$\begin{aligned} \psi(x, t) + \epsilon \frac{\partial\psi(x, t)}{\partial t} &= \exp\left(\frac{-i\epsilon V(x)}{\hbar}\right) \frac{(2\pi\hbar\epsilon i/m)^{1/2}}{A} \\ &\times \left[ \psi(x, t) + \frac{\hbar\epsilon i}{m} \frac{\partial^2\psi(x, t)}{\partial x^2} + \dots \right]. \end{aligned} \quad (27)$$

In order that both sides may agree to *zero* order in  $\epsilon$ , we must set

$$A = (2\pi\hbar\epsilon i/m)^{1/2}. \quad (28)$$

Then expanding the exponential containing  $V(x)$ , we get

$$\psi(x, t) + \epsilon \frac{\partial\psi}{\partial t} = \left(1 - \frac{i\epsilon}{\hbar} V(x)\right) \times \left(\psi(x, t) + \frac{\hbar\epsilon i}{2m} \frac{\partial^2\psi}{\partial x^2}\right). \quad (29)$$

---

<sup>13</sup>Really, these integrals are oscillatory and not defined, but they may be defined by using a convergence factor. Such a factor is automatically provided by  $\psi(x - \xi, t)$  in (24). If a more formal procedure is desired replace  $\hbar$  by  $\hbar(1 - i\delta)$ , for example, where  $\delta$  is a small positive number, and then let  $\delta \rightarrow 0$ .

Cancelling  $\psi(x, t)$  from both sides, and comparing terms to first order in  $\epsilon$  and multiplying by  $-\hbar/i$  one obtains

$$-\frac{\hbar}{i} \frac{\partial \psi}{\partial t} = \frac{1}{2m} \left( \frac{\hbar}{i} \frac{\partial}{\partial x} \right)^2 \psi + V(x) \psi, \quad (30)$$

which is Schrödinger's equation for the problem in question.

The equation for  $\chi^*$  can be developed in the same way, but adding a factor *decreases* the time by one step, i.e.,  $\chi^*$  satisfies an equation like (30) but with the sign of the time reversed. By taking complex conjugates we can conclude that  $\chi$  satisfies the same equation as  $\psi$ , i.e., an experiment can be defined by the particular state  $\chi$  to which it corresponds.<sup>14</sup>

This example shows that most of the contribution to  $\psi(x_{k+1}, t+\epsilon)$  comes from values of  $x_k$  in  $\psi(x_k, t)$  which are quite close to  $x_{k+1}$  (distant of order  $\epsilon^{1/2}$ ) so that the integral equation (23) can, in the limit, be replaced by a differential equation. The "velocities,"  $(x_{k+1} - x_k)/\epsilon$  which are important are very high, being of order  $(\hbar/m\epsilon)^{1/2}$  which diverges as  $\epsilon \rightarrow 0$ . The paths involved are, therefore, continuous but possess no derivative. They are of a type familiar from study of Brownian motion.

It is these large velocities which make it so necessary to be careful in approximating  $S(x_{k+1}, x_k)$  from Eq. (11).<sup>15</sup> To replace  $V(x_{k+1})$  by  $V(x_k)$  would, of course, change the exponent in (18) by  $i\epsilon[V(x_k) - V(x_{k+1})]/\hbar$  which is of order  $\epsilon(x_{k+1} - x_k)$ , and thus lead to unimportant terms of higher order than  $\epsilon$  on the right-hand side of (29). It is for this reason that (20) and (21) are equally satisfactory approximations to  $S(x_{i+1}, x_i)$  when there is no vector potential. A term, linear in velocity, however, arising from a vector potential, as  $A \dot{x} dt$  must be handled more carefully. Here a term in  $S(x_{k+1}, x_k)$  such as  $A(x_{k+1}) \times (x_{k+1} - x_k)$  differs from  $A(x_k)(x_{k+1} - x_k)$  by a term of order

---

<sup>14</sup>Dr. Hartland Snyder has pointed out to me, in private conversation, the very interesting possibility that there may be a generalization of quantum mechanics in which the states measured by experiment cannot be prepared; that is, there would be no state into which a system may be put for which a particular experiment gives certainty for a result. The class of functions  $\chi$  is not identical to the class of available states  $\psi$ . This would result if, for example,  $\chi$  satisfied a different equation than  $\psi$ .

<sup>15</sup>Equation (18) is actually exact when (11) is used for  $S(x_{i+1}, x_i)$  for arbitrary  $\epsilon$  for cases in which the potential does not involve  $x$  to higher powers than the second (e.g., free particle, harmonic oscillator). It is necessary, however, to use a more accurate value of  $A$ . One can define  $A$  in this way. Assume classical particles with  $k$  degrees of freedom start from the point  $x_i, t_i$ , with uniform density in momentum space. Write the number of particles having a given component of momentum in range  $dp$  as  $dp/p_0$ , with  $p_0$ , constant. Then  $A = (2\pi\hbar i/p_0)^{k/2} \rho^{-1/2}$ , where  $\rho$  is the density in  $k$  dimensional coordinate space  $x_{i+1}$  of these particles at time  $t_{i+1}$ .

$(x_{k+1} - x_k)^2$ , and, therefore, of order  $\epsilon$ . Such a term would lead to a change in the resulting wave equation. For this reason the approximation (21) is not a sufficiently accurate approximation to (11) and one like (20), (or (19) from which (20) differs by terms of order higher than  $\epsilon$ ) must be used. If  $\mathbf{A}$  represents the vector potential and  $\mathbf{p} = (\hbar/i)\nabla$ , the momentum operator, then (20) gives, in the Hamiltonian operator, a term  $(1/2m)(\mathbf{p} - (e/c\mathbf{A}) \cdot (\mathbf{p} - e/c)\mathbf{A})$ , while (21) gives  $(1/2m)(\mathbf{p} \cdot \mathbf{p} - (2e/c)\mathbf{A} \cdot \mathbf{p} + (e^2/c^2)\mathbf{A} \cdot \mathbf{A})$ . These two expressions differ by  $(\hbar e/2imc) \nabla \cdot \mathbf{A}$  which may not be zero. The question is still more important in the coefficient of terms which are quadratic in the velocities. In these terms (19) and (20) are not sufficiently accurate representations of (11) in general. It is when the coefficients are constant that (19) or (20) can be substituted for (11). If an expression such as (19) is used, say for spherical coordinates, when it is not a valid approximation to (11), one obtains a Schrödinger equation in which the Hamiltonian operator has some of the momentum operators and coordinates in the wrong order. Equation (11) then resolves the ambiguity in the usual rule to replace  $p$  and  $q$  by the non-commuting quantities  $(\hbar/i)(\partial/\partial q)$  and  $q$  in the classical Hamiltonian  $H(p, q)$ .

It is clear that the statement (11) is independent of the coordinate system. Therefore, to find the differential wave equation it gives in any coordinate system, the easiest procedure is first to find the equations in Cartesian coordinates and then to transform the coordinate system to the one desired. It suffices, therefore, to show the relation of the postulates and Schrödinger's equation in rectangular coordinates.

The derivation given here for one dimension can be extended directly to the case of three-dimensional Cartesian coordinates for any number,  $K$ , of particles interacting through potentials with one another, and in a magnetic field, described by a vector potential. The terms in the vector potential require completing the square in the exponent in the usual way for Gaussian integrals. The variable  $x$  must be replaced by the set  $x^{(1)}$  to  $x^{(3K)}$  where  $x^{(1)}, x^{(2)}, x^{(3)}$  are the coordinates of the first particle of mass  $m_1$ ,  $x^{(4)}, x^{(5)}, x^{(6)}$  of the second of mass  $m_2$ , etc. The symbol  $dx$  is replaced by  $dx^{(1)}dx^{(2)}\dots dx^{(3K)}$ , and the integration over  $dx$  is replaced by a  $3K$ -fold integral. The constant  $A$  has, in this case, the value  $A = (2\pi\hbar ei/m_1)^{1/2}(2\pi\hbar ei/m_2)^{1/2}\dots(2\pi\hbar ei/m_K)^{1/2}$ . The Lagrangian is the classical Lagrangian for the same problem, and the Schrödinger equation resulting will be that which corresponds to the classical Hamiltonian, derived from this Lagrangian. The equations in any other coordinate system may be obtained by transformation. Since this includes all cases for which Schrödinger's equation has been checked with experiment, we may say our

postulates are able to describe what can be described by non-relativistic quantum mechanics, neglecting spin.

## 7. Discussion of the Wave Equation

### The Classical Limit

This completes the demonstration of the equivalence of the new and old formulations. We should like to include in this section a few remarks about the important equation (18).

This equation gives the development of the wave function during a small time interval. It is easily interpreted physically as the expression of Huygens' principle for matter waves. In geometrical optics the rays in an inhomogeneous medium satisfy Fermat's principle of least *time*. We may state Huygens' principle in wave optics in this way: If the amplitude of the wave is known on a given surface, the amplitude at a near by point can be considered as a sum of contributions from all points of the surface. Each contribution is delayed in phase by an amount proportional to the *time* it would take the light to get from the surface to the point along the ray of least *time* of geometrical optics. We can consider (22) in an analogous manner starting with Hamilton's first principle of least *action* for classical or "geometrical" mechanics. If the amplitude of the wave  $\psi$  is known on a given "surface," in particular the "surface" consisting of all  $x$  at time  $t$ , its value at a particular nearby point at time  $t + \epsilon$ , is a sum of contributions from all points of the surface at  $t$ . Each contribution is delayed in phase by an amount proportional to the *action* it would require to get from the surface to the point along the path of least action of classical mechanics. <sup>16</sup>

Actually Huygens' principle is not correct in optics. It is replaced by Kirchhoff's modification which requires that both the amplitude and its derivative must be known on the adjacent surface. This is a consequence of the fact that the wave equation in optics is second order in the time. The wave equation of quantum mechanics is first order in the time; therefore, Huygens' principle *is* correct for matter waves, action replacing time.

The equation can also be compared mathematically to quantities appearing in the usual formulations. In Schroedinger's method the development of the wave function with time is given by

$$-\frac{\hbar}{i} \frac{\partial \psi}{\partial t} = \mathbf{H}\psi, \quad (31)$$

---

<sup>16</sup>See in this connection the very interesting remarks of Schroedinger, Ann. d. Physik **79**, 489 (1926).

which has the solution (for any  $\epsilon$  if  $\mathbf{H}$  is time independent)

$$\psi(x, t + \epsilon) = \exp(-i\epsilon\mathbf{H}/\hbar)\psi(x, t). \quad (32)$$

Therefore, Eq. (18) expresses the operator  $\exp(-i\epsilon\mathbf{H}/\hbar)$  by an approximate integral operator for small  $\epsilon$ .

From the point of view of Heisenberg one considers the position at time  $t$ , for example, as an operator  $\mathbf{x}$ . The position  $\mathbf{x}'$  at a later time  $t + \epsilon$  can be expressed in terms of that at time  $t$  by the operator equation

$$\mathbf{x}' = \exp(i\epsilon\mathbf{H}/\hbar)\mathbf{x}\exp(-i\epsilon\mathbf{H}/\hbar). \quad (33)$$

The transformation theory of Dirac allows us to consider the wave function at time  $t + \epsilon$ ,  $\psi(x', t + \epsilon)$ , as representing a state in a representation in which  $\mathbf{x}'$  is diagonal, while  $\psi(x, t)$  represents the same state in a representation in which  $\mathbf{x}$  is diagonal. They are, therefore, related through the transformation function  $(x'|x)_\epsilon$ , which relates these representations:

$$\psi(x', t + \epsilon) = \int (x'|x)_\epsilon \psi(x, t) dx.$$

Therefore, the content of Eq. (18) is to show that for small  $\epsilon$  we can set

$$(x'|x)_\epsilon = (1/A)\exp(iS(x', x)/\hbar) \quad (34)$$

with  $S(x', x)$  defined as in (11).

The close analogy between  $(x'|x)_\epsilon$  and the quantity  $\exp(iS(x', x)/\hbar)$  has been pointed out on several occasions by Dirac.<sup>17</sup> In fact, we now see that to sufficient approximations the two quantities may be taken to be proportional to each other. Dirac's remarks were the starting point of the present development. The points he makes concerning the passage to the classical limit  $\hbar \rightarrow 0$  are very beautiful, and I may perhaps be excused for briefly reviewing them here.

First we note that the wave function at  $x''$  at time  $t''$  can be obtained from that at  $x'$  at time  $t'$  by

$$\begin{aligned} \psi(x'', t'') &= \lim_{\epsilon \rightarrow 0} \int \dots \int \times \\ &\times \exp \left[ \frac{i}{\hbar} \sum_{i=0}^{j-1} S(x_{i+1}, x_i) \right] \times \psi(x', t') \frac{dx_0}{A} \frac{dx_1}{A} \dots \frac{dx_{j-1}}{A}, \end{aligned} \quad (35)$$

---

<sup>17</sup>P. A. M. Dirac, *The Principles of Quantum Mechanics* (The Clarendon Press, Oxford, 1935), second edition, Section 33; also, Physik. Zeits. Sowjetunion **3**, 64 (1933).

where we put  $x_0 \equiv x''$  and  $x_j \equiv x''$  where  $i\epsilon = t'' - t'$  (between the times  $t'$  and  $t''$  we assume no restriction is being put on the region of integration). This can be seen either by repeated applications of (18) or directly from Eq. (15). Now we ask, as  $\hbar \rightarrow 0$  what values of the intermediate coordinates  $x_i$ , contribute most strongly to the integral? These will be the values most likely to be found by experiment and therefore will determine, in the limit, the classical path. If  $\hbar$  is very small, the exponent will be a very rapidly varying function of any of its variables  $x_i$ . As  $x_i$  varies, the positive and negative contributions of the exponent nearly cancel. The region at which  $x_i$ , contributes most strongly is that at which the phase of the exponent varies least rapidly with  $x_i$  (method of stationary phase). Call the sum in the exponent  $S$ ;

$$S = \sum_{i=0}^{j-1} S(x_{i+1}, x_i). \quad (36)$$

Then the classical orbit passes, approximately, through those points  $x_i$  at which the rate of change of  $S$  with  $x_i$ , is small, or in the limit of small  $\hbar$ , zero, i.e., the classical orbit passes through the points at which  $\partial S / \partial x_i$  for all  $x_i$ . Taking the limit  $\epsilon \rightarrow 0$ , (36) becomes in view of (11)

$$S = \int_{t'}^{t''} L(\dot{x}(t), x(t)) dt. \quad (37)$$

We see then that the classical path is that for which the integral (37) suffers no first-order change on varying the path. This is Hamilton's principle and leads directly to the Lagrangian equations of motion.

## 8. Operator Algebra

### Matrix Elements

Given the wave function and Schroedinger's equation, of course all of the machinery of operator or matrix algebra can be developed. It is, however, rather interesting to express these concepts in a somewhat different language more closely related to that used in stating the postulates. Little will be gained by this in elucidating operator algebra. In fact, the results are simply a translation of simple operator equations into a somewhat more cumbersome notation. On the other hand, the new notation and point of view are very useful in certain applications described in the introduction. Furthermore, the form of the equations permits natural extension to a wider

class of operators than is usually considered (e.g., ones involving quantities referring to two or more different times). If any generalization to a wider class of action functionals is possible, the formulae to be developed will play an important role.

We discuss these points in the next three sections. This section is concerned mainly with definitions. We shall define a quantity which we call a transition element between two states. It is essentially a matrix element. But instead of being the matrix element between a state  $\psi$  and another  $\chi$  corresponding to the *same* time, these two states will refer to different times. In the following section a fundamental relation between transition elements will be developed from which the usual commutation rules between coordinate and momentum may be deduced. The same relation also yields Newton's equation of motion in matrix form. Finally, in Section 10 we discuss the relation of the Hamiltonian to the operation of displacement in time.

We begin by defining a transition element in terms of the probability of transition from one state to another. More precisely, suppose we have a situation similar to that described in deriving (17). The region  $R$  consists of a region  $R'$  previous to  $t'$ , all space between  $t'$  and  $t''$  and the region  $R''$  after  $t''$ . We shall study the probability that a system in region  $R'$  is later found in region  $R''$ . This is given by (17). We shall discuss in this section how it changes with changes in the form of the Lagrangian between  $t'$  and  $t''$ . In Section 10 we discuss how it changes with changes in the preparation  $R'$  or the experiment  $R''$ .

The state at time  $t'$  is defined completely by the preparation  $R'$ . It can be specified by a wave function  $\psi(x', t')$  obtained as in (15), but containing only integrals up to the time  $t'$ . Likewise, the state characteristic of the experiment (region  $R''$ ) can be defined by a function  $\chi(x'', t'')$  obtained from (16) with integrals only beyond  $t''$ . The wave function  $\psi(x'', t'')$  at time  $t''$  can, of course, also be gotten by appropriate use of (15). It can also be gotten from  $\psi(x', t')$  by (35). According to (17) with  $t''$  used instead of  $t$ , the probability of being found in  $\chi$  it prepared in  $\psi$  is the square of what we shall call the transition amplitude  $\int \chi^*(x'', t'')\psi(x'', t'')dx''$ . We wish to express this in terms of  $\chi$  at  $t''$  and  $\psi$  at  $t'$ . This we can do with the aid of (35). Thus, the chance that a system prepared in state  $\psi_{t'}$  at time  $t'$  will be found after  $t''$  to be in a state  $\chi_{t''}$  is the square of the transition amplitude

$$\begin{aligned} \langle \chi_{t''} | 1 | \psi_{t'} \rangle_S = \lim_{\epsilon \rightarrow 0} & \int \dots \int \chi^*(x'', t'') \times \\ & \times \exp(iS/\hbar) \psi(x', t') \frac{dx_0}{A} \dots \frac{dx_{j-1}}{A} dx_j, \end{aligned} \quad (38)$$

where we have used the abbreviation (36).

In the language of ordinary quantum mechanics if the Hamiltonian,  $\mathbf{H}$ , is constant,  $\psi(x, t'') = \exp[-i(t'' - t')\mathbf{H}/\hbar]\psi(x, t')$  so that (38) is the matrix element of  $\exp[-i(t'' - t')\mathbf{H}/\hbar]$  between states  $\chi_{t''}$  and  $\psi_{t'}$ .

If  $F$  is any function of the coordinates  $x_i$  for  $t' < t_i < t''$ , we shall define the transition element of  $F$  between the states  $\psi$  at  $t'$  and  $\chi$  at  $t''$  for the action  $S$  as ( $x'' \equiv x_j$ ,  $x' \equiv x_0$ ):

$$\begin{aligned} \langle \chi_{t''} | F | \psi_{t'} \rangle = \lim_{\epsilon \rightarrow 0} & \int \dots \int \times \chi^*(x'', t'') F(x_0, x_1, \dots x_i) \cdot \\ & \cdot \exp \left[ \frac{i}{\hbar} \sum_{i=0}^{j-1} S(x_{i+1}, x_i) \right] \psi(x', t') \frac{dx_0}{A} \dots \frac{dx_{j-1}}{A} dx_i. \end{aligned} \quad (39)$$

In the limit  $\epsilon \rightarrow 0$ ,  $F$  is a functional of the path  $x(t)$ .

We shall see presently why such quantities are important. It will be easier to understand if we stop for a moment to find out what the quantities correspond to in conventional notation. Suppose  $F$  is simply  $x_k$ , where  $k$  corresponds to some time  $t = t_k$ . Then on the right-hand side of (39) the integrals from  $x_0$  to  $x_{k-1}$  may be performed to produce  $\psi(x_k, t)$  or  $\exp[-i(t - t')\mathbf{H}/\hbar]\psi_{t'}$ . In like manner the integrals on  $x_i$  for  $j \geq i > k$  give  $\chi^*(x_k, t)$  or  $\{\exp[-i(t'' - t)\mathbf{H}/\hbar]\chi_{t''}\}$ . Thus, the transition element of  $x_k$ ,

$$\begin{aligned} \langle \chi_{t''} | F | \psi_{t'} \rangle_S = & \int \chi_{t''}^* e^{(i/\hbar)\mathbf{h}(t'' - t)} x e^{-(i/\hbar)\mathbf{H}(t - t')} \psi_{t'} dx = \\ & = \int \chi^*(x, t) x \psi(x, t) dx \end{aligned} \quad (40)$$

is the matrix element of  $\mathbf{x}$  at time  $t = t_k$  between the state which would develop at time  $t$  from  $\psi_{t'}$  at  $t'$  and the state which will develop from time  $t$  to  $\chi_{t''}$  at  $t''$ . It is, therefore, the matrix element of  $\mathbf{x}(t)$  between these states.

Likewise, according to (39) with  $F = x_{k+1}$ , the transition element of  $x_{k+1}$  is the matrix element of  $\mathbf{x}(t + \epsilon)$ . The transition element of  $F = (x_{k+1} - x_k)/\epsilon$  is the matrix element of  $(\mathbf{x}(t + \epsilon) - \mathbf{x}(t))/\epsilon$  or of  $i(\mathbf{Hx} - \mathbf{xH})/\hbar$ , as is easily shown from (40). We can call this the matrix element of velocity  $\dot{x}(t)$ .

Suppose we consider a second problem which differs from the first because, for example, the potential is augmented by a small amount  $U(, \mathbf{xt})$ . Then in the new problem the quantity replacing  $S$  is  $S' = S + \sum_i \epsilon U(x_i, t_i)$ . Substitution into (38) leads directly to

$$\langle \chi_{t''} | 1 | \psi_{t'} \rangle_{S'} = \left\langle \chi_{t''} \left| \exp \frac{i\epsilon}{\hbar} \sum_{i=1}^j U(x_i, t_i) \right| \psi_{t'} \right\rangle_S. \quad (41)$$

Thus, transition elements such as (39) are important insofar as  $F$  may arise in some way from a change  $\delta S$  in an action expression. We denote, by observable functionals, those functionals  $F$  which can be defined, (possibly indirectly) in terms of the changes which are produced by possible changes in the action  $S$ . The condition that a functional be observable is somewhat similar to the condition that an operator be Hermitian. The observable functionals are a restricted class because the action must remain a quadratic function of velocities. From one observable functional others may be derived, for example, by

$$\langle \chi_{t''} | F | \psi_{t'} \rangle_{S'} = \left\langle \chi_{t''} \left| F \exp \frac{i\epsilon}{\hbar} \sum_{i=1}^j U(x_i, t_i) \right| \psi_{t'} \right\rangle_S \quad (42)$$

which is obtained from (39).

Incidentally, (41) leads directly to an important perturbation formula. If the effect of  $U$  is small the exponential can be expanded to first order in  $U$  and we find

$$\langle \chi_{t''} | 1 | \psi_{t'} \rangle_{S'} = \left\langle \chi_{t''} | 1 | \psi_{t'} \right\rangle_S + \frac{i}{\hbar} \langle \chi_{t''} | \sum_i \epsilon U(x_i, t_i) | \psi_{t'} \rangle. \quad (43)$$

Of particular importance is the case that  $\chi_{t''}$  is a state in which  $\psi_{t'}$  would not be found at all were it not for the disturbance,  $U$  (i.e.,  $\langle \chi_{t''} | 1 | \psi_{t'} \rangle_S = 0$ ) Then

$$\frac{1}{\hbar^2} |\langle \chi_{t''} | \sum_i \epsilon U(x_i, t_i) | \psi_{t'} \rangle_S|^2 \quad (44)$$

is the probability of transition as induced to first order by the perturbation. In ordinary notation,

$$\langle \chi_{t''} | \sum_i \epsilon U(x_i, t_i) | \psi_{t'} \rangle_S = \int \left\{ \int \chi_{t''}^* e^{-(i/\hbar)\mathbf{H}(t''-t)} \mathbf{U} e^{-(i/\hbar)\mathbf{H}(t-t')} \psi_{t'} dx \right\} dt$$

so that (44) reduces to the usual expression <sup>18</sup> for time dependent perturbations.

## 9. Newton's Equations

### The Commutation Relation

In this section we find that different functionals may give identical results when taken between any two states. This equivalence between functionals

---

<sup>18</sup>P. A. M. Dirac, *The Principles of Quantum Mechanics* (The Clarendon Press, Oxford, 1935), second edition, Section 47, Eq. (20)

is the statement of operator equations in the new language.

If  $F$  depends on the various coordinates, we can, of course, define a new functional  $\partial F/\partial x_k$  by differentiating it with respect to one of its variables, say  $x_k$  ( $0 < k < j$ ). If we calculate  $\langle \chi_{t''} | \partial F / \partial x_k | \psi_{t'} \rangle_S$  by (39) the integral on the right-hand side will contain  $\partial F / \partial x_k$ . The only other place that the variable  $x_k$  appears is in  $S$ . Thus, the integration on  $x_k$  can be performed by parts. The integrated part vanishes (assuming wave functions vanish at infinity) and we are left with the quantity  $-F(\partial/\partial x_k)\exp(iS/\hbar)$  in the integral. However,  $(\partial/\partial x_k)\exp(iS/\hbar) = (i/\hbar)(\partial S/\partial x_k)\exp(iS/\hbar)$ , so the right side represents the transition element of  $-(i/\hbar)F(\partial S/\partial x_k)$ , i.e.,

$$\left\langle \chi_{t''} \left| \frac{\partial F}{\partial x_k} \right| \psi_{t'} \right\rangle_S = -\frac{i}{\hbar} \left\langle \chi_{t''} \left| F \frac{\partial S}{\partial x_k} \right| \psi_{t'} \right\rangle_S. \quad (45)$$

This very important relation shows that two different functionals may give the same result for the transition element between any two states. We say they are equivalent and symbolize the relation by

$$-\frac{\hbar}{i} \frac{\partial F}{\partial x_k} \underset{S}{\leftrightarrow} F \frac{\partial S}{\partial x_k}, \quad (46)$$

the symbol  $\underset{S}{\leftrightarrow}$  emphasizing the fact that functionals equivalent under one action may not be equivalent under another. The quantities in (46) need not be observable. The equivalence is, nevertheless, true. Making use of (36) one can write

$$-\frac{\hbar}{i} \frac{\partial F}{\partial x_k} \underset{S}{\leftrightarrow} F \left[ \frac{\partial S(x_{k+1}, x_k)}{\partial x_k} + \frac{\partial S(x_k, x_{k-1})}{\partial x_k} \right]. \quad (47)$$

This equation is true to zero and first order in  $\epsilon$  and has as consequences the commutation relations of momentum and coordinate, as well as the Newtonian equations of motion in matrix form.

In the case of our simple one-dimensional problem,  $S(x_{i+1}, x_i)$  is given by the expression (15), so that

$$\partial S(x_{k+1}, x_k) / \partial x_k = -m(x_{k+1} - x_k) / \epsilon,$$

and

$$\partial S(x_k, x_{k-1}) / \partial x_k = +m(x_k - x_{k-1}) / \epsilon - \epsilon V'(x_k);$$

where we write  $V'(x)$  for the derivative of the potential, or force. Then (47) becomes

$$-\frac{\hbar}{i} \frac{\partial F}{\partial x_k} \underset{S}{\leftrightarrow} F \left[ -m \left( \frac{x_{k+1} - x_k}{\epsilon} - \frac{x_k - x_{k-1}}{\epsilon} \right) - \epsilon V'(x_k) \right]. \quad (48)$$

If  $F$  does not depend on the variable  $x_k$ , this gives Newton's equations of motion. For example, if  $F$  is constant, say unity, (48) just gives (dividing by  $\epsilon$ )

$$0 \underset{S}{\leftrightarrow} -\frac{m}{\epsilon} \left( \frac{x_{k+1} - x_k}{\epsilon} - \frac{x_k - x_{k-1}}{\epsilon} \right) - V'(x_k).$$

Thus, the transition element of mass times acceleration  $[(x_{k+1} - x_k)/\epsilon - (x_k - x_{k-1})/\epsilon]/\epsilon$  between any two states is equal to the transition element of force  $-V'(x_k)$  between the same states. This is the matrix expression of Newton's law which holds in quantum mechanics.

What happens if  $F$  does depend upon  $x_k$ ? For example, let  $F = x_k$ . Then (48) gives, since  $\partial F / \partial x_k = 1$ ,

$$-\frac{\hbar}{i} \underset{S}{\leftrightarrow} x_k \left[ -m \left( \frac{x_{k+1} - x_k}{\epsilon} - \frac{x_k - x_{k-1}}{\epsilon} \right) - \epsilon V'(x_k) \right]$$

or, neglecting terms of order  $\epsilon$ ,

$$m \left( \frac{x_{k+1} - x_k}{\epsilon} \right) x_k - m \left( \frac{x_k - x_{k-1}}{\epsilon} \right) x_k \underset{S}{\leftrightarrow} \frac{\hbar}{i}. \quad (49)$$

In order to transfer an equation such as (49) into conventional notation, we shall have to discover what matrix corresponds to a quantity such as  $x_k x_{k+1}$ . It is clear from a study of (39) that if  $F$  is set equal to, say,  $f(x_k)g(x_{k+1})$ , the corresponding operator in (40) is

$$e^{-(i/\hbar)(t''-t-\epsilon)\mathbf{H}} g(\mathbf{x}) e^{-(i/\hbar)\epsilon\mathbf{H}} f(\mathbf{x}) e^{-(i/\hbar)(t-t')\mathbf{H}},$$

the matrix element being taken between the states  $\chi_{t''}$  and  $\psi_{t'}$ . The operators corresponding to functions of  $x_{k+1}$  will appear to the left of the operators corresponding to functions of  $x_k$ , i.e., *the order of terms in a matrix operator product corresponds to an order in time of the corresponding factors in a functional*. Thus, if the functional can and is written in such a way that in each term factors corresponding to later times appear to the left of factors corresponding to earlier terms, the corresponding operator can immediately be written down if the order of the operators is kept the same as in the functional.<sup>19</sup> Obviously, the order of factors in a functional is of no consequence. The ordering just facilitates translation into conventional operator notation. To write Eq. (49) in the way desired for easy translation

---

<sup>19</sup>Dirac has also studied operators containing quantities referring to different times. See reference 2.

would require the factors in the second term on the left to be reversed in order. We see, therefore, that it corresponds to

$$\mathbf{p}\mathbf{x} - \mathbf{x}\mathbf{p} = \hbar/i$$

where we have written  $\mathbf{p}$  for the operator  $m\dot{\mathbf{x}}$ .

The relation between functionals and the corresponding operators is defined above in terms of the order of the factors in time. It should be remarked that this rule must be especially carefully adhered to when quantities involving velocities or higher derivatives are involved. The correct functional to represent the operator  $(\dot{x})^2$  is actually  $(x_{k+1} - x_k)/\epsilon(x_k - x_{k-1})/\epsilon$  rather than  $[(x_{k+1} - x_k)/\epsilon]^2$ . The latter quantity diverges as  $1/\epsilon$  as  $\epsilon \rightarrow 0$ . This may be seen by replacing the second term in (49) by its value  $x_{k+1} \cdot m(x_{k+1} - x_k)/\epsilon$  calculated an instant  $\epsilon$  later in time. This does not change the equation to zero order in  $\epsilon$ . We then obtain (dividing by  $\epsilon$ )

$$\left( \frac{x_{k+1} - x_k}{\epsilon} \right)^2 \underset{S}{\leftrightarrow} -\frac{\hbar}{im\epsilon}. \quad (50)$$

This gives the result expressed earlier that the root mean square of the “velocity”  $(x_{k+1} - x_k)/\epsilon$  between two successive positions of the path is of order  $\epsilon^{-1/2}$ .

It will not do then to write the functional for kinetic energy, say, simply as

$$\frac{1}{2}m[(x_{k+1} - x_k)/\epsilon]^2 \quad (51)$$

for this quantity is infinite as  $\epsilon \rightarrow 0$ . In fact, it is not an observable functional.

One can obtain the kinetic energy as an observable functional by considering the first-order change in transition amplitude occasioned by a change in the mass of the particle. Let  $m$  be changed to  $m(1 + \delta)$  for a short time, say  $\epsilon$ , around  $t_k$ . The change in the action is  $\frac{1}{2}\delta\epsilon m[(x_{k+1} - x_k)/\epsilon]^2$  the derivative of which gives an expression like (51). But the change in  $m$  changes the normalization constant  $1/A$  corresponding to  $dx_k$  as well as the action. The constant is changed from  $(2\pi\hbar\epsilon i/m)^{-1/2}$  to  $(2\pi\hbar\epsilon i/m(1 + \delta))^{-1/2}$  or by  $\frac{1}{2}\delta(2\pi\hbar\epsilon i/m)^{-1/2}$  to first order in  $\delta$ . The total effect of the change in mass in Eq. (38) to the first order in  $\delta$  is

$$\langle \chi_{t''} | \frac{1}{2}\delta\epsilon im[(x_{k+1} - x_k)/\epsilon]^2 / \hbar + \frac{1}{2}\delta|\psi_{t'}\rangle.$$

We expect the change of order  $\delta$  lasting for a time  $\epsilon$  to be of order  $\delta\epsilon$ . Hence, dividing by  $\delta\epsilon i/\hbar$ , we can define the kinetic energy functional as

$$\text{K.E.} = \frac{1}{2}m[(x_{k+1} - x_k)/\epsilon]^2 + \hbar/2\epsilon i. \quad (52)$$

This is finite as  $\epsilon \rightarrow 0$  in view of (50). By making use of an equation which results from substituting  $m(x_{k+1} - x_k)/\epsilon$  for  $F$  in (48) we can also show that the expression (52) is equal (to order  $\epsilon$ ) to

$$\text{K.E.} = \frac{1}{2}m \left( \frac{x_{k+1} - x_k}{\epsilon} \right) \left( \frac{x_k - x_{k-1}}{\epsilon} \right). \quad (53)$$

That is, the easiest way to produce observable functionals involving powers of the velocities is to replace these powers by a product of velocities, each factor of which is taken at a slightly different time.

## 10. The Hamiltonian

### Momentum

The Hamiltonian operator is of central importance in the usual formulation of quantum mechanics. We shall study in this section the functional corresponding to this operator. We could immediately define the Hamiltonian functional by adding the kinetic energy functional (52) or (53) to the potential energy. This method is artificial and does not exhibit the important relationship of the Hamiltonian to time. We shall define the Hamiltonian functional by the changes made in a state when it is displaced in time.

To do this we shall have to digress a moment to point out that the subdivision of time into *equal* intervals is not necessary. Clearly, any subdivision into instants  $t_i$  will be satisfactory; the limits are to be taken as the largest spacing,  $t_{i+1} - t_i$  approaches zero. The total action  $S$  must now be represented as a sum

$$S = \sum_i S(x_{i+1}, t_{i+1}; x_i, t_i), \quad (54)$$

where

$$S(x_{i+1}, t_{i+1}; x_i, t_i) = \int_{t_i}^{t_{i+1}} L(\dot{x}(t), x(t)) dt, \quad (55)$$

the integral being taken along the classical path between  $x_i$ , at  $t_i$  and  $x_{i+1}$  at  $t_{i+1}$ . For the simple one-dimensional example this becomes, with sufficient accuracy,

$$S(x_{i+1}, t_{i+1}; x_i, t_i) = \left\{ \frac{m}{2} \left( \frac{x_{i+1} - x_i}{t_{i+1} - t_i} \right)^2 - V(x_{i+1}) \right\} (t_{i+1} - t_i); \quad (56)$$

the corresponding normalization constant for integration on  $dx_i$  is  $A = (2\pi\hbar i(t_{i+1} - t_i)/m)^{-1/2}$ .

The relation of  $H$  to the change in a state with displacement in time can now be studied. Consider a state  $\psi(t)$  defined by a space-time region  $R'$ . Now imagine that we consider another state at time  $t$ ,  $\psi_\delta(t)$ , denoted by another region  $R'_\delta$ . Suppose the region  $R'_\delta$  is exactly the same as  $R'$  except that it is earlier by a time  $\delta$ , i.e., displaced bodily toward the past by a time  $\delta$ . All the apparatus to prepare the system for  $R'_\delta$  is identical to that for  $R'$  but is operated a time  $\delta$  sooner. If  $L$  depends explicitly on time, it, too, is to be displaced, i.e., the state  $\psi_\delta$  is obtained from the  $L$  used for state  $\psi$  except that the time  $t$  in  $L_\delta$  is replaced by  $t + \delta$ . We ask how does the state  $\psi_\delta$  differ from  $\psi$ ? In any measurement the chance of finding the system in a fixed region  $R''$  is different for  $R'$  and  $R'_\delta$ . Consider the change in the transition element  $\langle \chi | 1 | \psi_\delta \rangle_{S_\delta}$  produced by the shift  $\delta$ . We can consider this shift as effected by decreasing all values of  $t_i$  by  $\delta$  for  $i \leq k$  and leaving all  $t_i$  fixed for  $i > k$ , where the time  $t$  lies in the interval between  $t_{k+1}$  and  $t_k$ .<sup>20</sup> This change will have no effect on  $S(x_{i+1}, t_{i+1}; x_i, t_i)$  as defined by (55) as long as both  $t_{i+1}$  and  $t_i$  are changed by the same amount. On the other hand,  $S(x_{k+1}, t_{k+1}; x_k, t_k)$  is changed to  $S(x_{k+1}, t_{k+1}; x_k, t_k - \delta)$ . The constant  $1/A$  for the integration on  $dx_k$ , is also altered to  $(2\pi\hbar i(t_{k+1} - t_k + \delta)/m)^{-1/2}$ . The effect of these changes on the transition element is given to the first order in  $\delta$  by

$$\langle \chi | 1 | \psi \rangle_S - \langle \chi | 1 | \psi_\delta \rangle_{S_\delta} = \frac{i\delta}{\hbar} \langle \chi | H_k | \psi \rangle_S, \quad (57)$$

here the Hamiltonian functional  $H_k$  is defined by

$$H_k = \frac{\partial S(x_{k+1}, t_{k+1}; x_k, t_k)}{\partial t_k} + \frac{\hbar}{2i(t_{k+1} - t_k)}. \quad (58)$$

---

<sup>20</sup>From the point of view of mathematical rigor, if  $\delta$  is finite, as  $\epsilon \rightarrow 0$  one gets into difficulty in that, for example, the interval  $t_{k+1} - t_k$  is kept finite. This can be straightened out by assuming  $\delta$  to vary with time and to be turned on smoothly before  $t = t_k$  and turned off smoothly after  $t = t_k$ . Then keeping the time variation of  $\delta$  fixed, let  $\epsilon \rightarrow 0$ . Then seek the first-order change as  $\delta \rightarrow 0$ . The result is essentially the same as that of the crude procedure used above.

The last term is due to the change in  $1/A$  and serves to keep  $H_k$  finite as  $\epsilon \rightarrow 0$ . For example, for the expression (56) this becomes

$$H_k = \frac{m}{2} \left( \frac{x_{k+1} - x_k}{t_{k+1} - t_k} \right) + \frac{\hbar}{2i(t_{k+1} - t_k)} + V(x_{k+1}),$$

which is just the sum of the kinetic energy functional (52) and that of the potential energy  $V(x_{k+1})$ .

The wave function  $\psi_\delta(x, t)$  represents, of course, the same state as  $\psi(x, t)$  will be after time  $\delta$ , i.e.,  $\psi(x, t + \delta)$ . Hence, (57) is intimately related to the operator equation (31).

One could also consider changes occasioned by a time shift in the final state  $\chi$ . Of course, nothing new results in this way for it is only the relative shift of  $\chi$  and  $\psi$  which counts. One obtains an alternative expression

$$H_k = -\frac{\partial S(x_{k+1}, t_{k+1}; x_k, t_k)}{\partial t_{k+1}} + \frac{\hbar}{2i(t_{k+1} - t_k)}. \quad (59)$$

This differs from (58) only by terms of order  $\epsilon$ . The time rate of change of a functional can be computed by considering the effect of shifting both initial and final state together. This has the same effect as calculating the transition element of the functional referring to a later time. What results is the analog of the operator equation

$$\frac{\hbar}{i} \dot{\mathbf{f}} = \mathbf{H}\mathbf{f} - \mathbf{f}\mathbf{H}.$$

The momentum functional  $p_t$  can be defined in an analogous way by considering the changes made by displacements of position:

$$\langle \chi | 1 | \psi \rangle_S - \langle \chi | 1 | \psi_\Delta \rangle_{S\Delta} = \frac{i\Delta}{\hbar} \langle \chi | p_k | \psi \rangle_S.$$

The state  $\psi_\Delta$  is prepared from a region  $R'_\Delta$  which is identical to region  $R'$  except that it is moved a distance  $\Delta$  in space. (The Lagrangian, if it depends explicitly on  $x$ , must be altered to  $L_\Delta = L(\dot{x}, x - \Delta)$  for times previous to  $t$ .) One finds <sup>21</sup>

$$p_k = \frac{\partial S(x_{k+1}, x_k)}{\partial x_{k+1}} = -\frac{\partial S(x_{k+1}, x_k)}{\partial x_k}. \quad (60)$$

---

<sup>21</sup>We did not immediately substitute  $p_i$  from (60) into (47) because (47) would then no longer have been valid to both zero order and the first order in  $\epsilon$ . We could derive the commutation relations, but not the equations of motion. The two expressions in (60) represent the momenta at each end of the interval  $t_i$ , to  $t_{i+1}$ . They differ by  $\epsilon V'(x_{k+1})$  because of the force acting during the time  $\epsilon$

Since  $\psi_\Delta(x, t)$  is equal to  $\psi(x - \Delta, t)$ , the close connection between  $p_k$  and the  $x$ -derivative of the wave function is established.

Angular momentum operators are related in an analogous way to rotations.

The derivative with respect to  $t_{i+1}$  of  $S(x_{i+1}, t_{i+1}; x_i, t_i)$  appears in the definition of  $H_i$ . The derivative with respect to  $t_{i+1}$  defines  $p_i$ . But the derivative with respect to  $t_{i+1}$  of  $S(x_{i+1}, t_{i+1}; x_i, t_i)$  is related to the derivative with respect to  $x_{i+1}$ , for the function  $S(x_{i+1}, t_{i+1}; x_i, t_i)$  defined by (55) satisfies the Hamilton-Jacobi equation. Thus, the Hamilton-Jacobi equation is an equation expressing  $H_i$ , in terms of the  $p_i$ . In other words, it expresses the fact that time displacements of states are related to space displacements of the same states. This idea leads directly to a derivation of the Schrödinger equation which is far more elegant than the one exhibited in deriving Eq. (30).

## 11. Inadequacies of the Formulation

The formulation given here suffers from a serious drawback. The mathematical concepts needed are new. At present, it requires an unnatural and cumbersome subdivision of the time interval to make the meaning of the equations clear. Considerable improvement can be made through the use of the notation and concepts of the mathematics of functionals. However, it was thought best to avoid this in a first presentation. One needs, in addition, an appropriate measure for the space of the argument functions  $x(t)$  of the functionals.<sup>22</sup>

It is also incomplete from the physical standpoint. One of the most important characteristics of quantum mechanics is its invariance under unitary transformations. These correspond to the canonical transformations of classical mechanics. Of course, the present formulation, being equivalent to ordinary formulations, can be mathematically demonstrated to be invariant under these transformations. However, it has not been formulated in such a way that it is *physically* obvious that it is invariant. This incompleteness shows itself in a definite way. No direct procedure has been outlined to

---

<sup>22</sup>There are very interesting mathematical problems involved in the attempt to avoid the subdivision and limiting processes. Some sort of complex measure is being associated with the space of functions  $x(t)$ . Finite results can be obtained under unexpected circumstances because the measure is not positive everywhere, but the contributions from most of the paths largely cancel out. These curious mathematical problems are sidestepped by the subdivision process. However, one feels as Cavalieri must have felt calculating the volume of a pyramid before the invention of calculus.

describe measurements of quantities other than position. Measurements of momentum, for example, of one particle, can be defined in terms of measurements of positions of other particles. The result of the analysis of such a situation does show the connection of momentum measurements to the Fourier transform of the wave function. But this is a rather roundabout method to obtain such an important physical result. It is to be expected that the postulates can be generalized by the replacement of the idea of “paths in a region of space-time  $R$ ” to “paths of class  $R$ ,” or “paths having property  $R$ .” But which properties correspond to which physical measurements has not been formulated in a general way.

## 12. A Possibility Generalization

The formulation suggests an obvious generalization. There are interesting classical problems which satisfy a principle of least action but for which the action cannot be written as an integral of a function of positions and velocities. The action may involve accelerations, for example. Or, again, if interactions are not instantaneous, it may involve the product of coordinates at two different times, such as  $\int x(t)x(t+T)dt$ . The action, then, cannot be broken up into a sum of small contributions as in (10). As a consequence, no wave function is available to describe a state. Nevertheless, a transition probability can be defined for getting from a region  $R'$  into another  $R''$ . Most of the theory of the transition elements  $\langle \chi_{t''}|F|\psi_t \rangle_S$  can be carried over. One simply invents a symbol, such as  $\langle R''|F|R' \rangle_S$  by an equation such as (39) but with the expressions (19) and (20) for  $\psi$  and  $\chi$  substituted, and the more general action substituted for  $S$ . Hamiltonian and momentum functionals can be defined as in section (10). Further details may be found in a thesis by the author.<sup>23</sup>

## 13. Application to Eliminate Field Oscillators

One characteristic of the present formulation is that it can give one a sort of bird’s-eye view of the space-time relationships in a given situation. Before

---

<sup>23</sup>The theory of electromagnetism described by J. A. Wheeler and R. P. Feynman, Rev. Mod. Phys. **17**, 157 (1945) can be expressed in a principle of least action involving the coordinates of particles alone. It was an attempt to quantize this theory, without reference to the fields, which led the author to study the formulation of quantum mechanics given here. The extension of the ideas to cover the case of more general action functions was developed in his Ph.D. thesis, “The principle of least action in quantum mechanics” submitted to Princeton University, 1942.

the integrations on the  $x$ , are performed in an expression such as (39) one has a sort of format into which various  $F$  functionals may be inserted. One can study how what goes on in the quantum-mechanical system at different times is interrelated. To make these vague remarks somewhat more definite, we discuss an example.

In classical electrodynamics the fields describing, for instance, the interaction of two particles can be represented as a set of oscillators. The equations of motion of these oscillators may be solved and the oscillators essentially eliminated (Lienard and Wiechert potentials). The interactions which result involve relationships of the motion of one particle at one time, and of the other particle at another time. In quantum electrodynamics the field is again represented as a set of oscillators. But the motion of the oscillators cannot be worked out and the oscillators eliminated. It is true that the oscillators representing longitudinal waves may be eliminated. The result is instantaneous electrostatic interaction. The electrostatic elimination is very instructive as it shows up the difficulty of self-interaction very distinctly. In fact, it shows it up so clearly that there is no ambiguity in deciding what term is incorrect and should be omitted. This entire process is not relativistically invariant, nor is the omitted term. It would seem to be very desirable if the oscillators, representing transverse waves, could also be eliminated. This presents an almost insurmountable problem in the conventional quantum mechanics. We expect that the motion of a particle  $a$  at one time depends upon the motion of  $b$  at a previous time, and *vice versa*. A wave function  $\psi(x_a, x_b; t)$ , however, can only describe the behavior of both particles at one time. There is no way to keep track of what  $b$  did in the past in order to determine the behavior of  $a$ . The only way is to specify the state of the set of oscillators at  $t$ , which serve to "remember" what  $b$  (and  $a$ ) had been doing.

The present formulation permits the solution of the motion of all the oscillators and their complete elimination from the equations describing the particles. This is easily done. One must simply solve for the motion of the oscillators before one integrates over the various variables  $x_i$ , for the particles. It is the integration over  $x_i$  which tries to condense the past history into a single state function. This we wish to avoid. Of course, the result depends upon the initial and final states of the oscillator. If they are specified, the result is an equation for  $\langle \chi_{t''} | 1 | \psi_t \rangle$  like (38), but containing as a factor, besides  $\exp(iS/\hbar)$  another functional  $G$  depending only on the coordinates describing the paths of the particles.

We illustrate briefly how this is done in a very simple case. Suppose a particle, coordinate  $x(t)$ , Lagrangian  $L(\dot{x}, x)$  interacts with an oscillator,

coordinate  $g(t)$ , Lagrangian  $\frac{1}{2}(\dot{q}^2 - \omega^2 q^2)$  through a term  $\gamma(x, t)q(t)$  in the Lagrangian for the system. Here  $\gamma(x, t)$  is any function of the coordinate  $x(t)$  of the particle and the time.<sup>24</sup> Suppose we desire the probability of a transition from a state at time  $t'$ , in which the particle's wave function is  $\psi_{t'}$  and the oscillator is in energy level  $n$ , to a state at  $t''$  with the particle in  $\chi_{t''}$  and oscillator in level  $m$ . This is the square of

$$\langle \chi_{t''} \varphi_m | 1 | \psi_{t'} \varphi_n \rangle_{S_p + S_0 + S_I} = \int \dots \int \varphi_m^*(q_i) \chi_{t''}^*(x_i) \\ \times \exp \frac{i}{\hbar} (S_p + S_0 + S_1) \psi_{t'}(x_0) \varphi_n(q_0) \cdot \frac{dx_0}{A} \frac{dq_0}{a} \dots \frac{dx_{j-1}}{A} \frac{dq_{j-1}}{a} dx_i dq_i. \quad (61)$$

Here  $\varphi_n(q)$  is the wave function for the oscillator in state  $n$ ,  $S_p$  is the action

$$\sum_{i=0}^{j-1} S_p(x_{i+1}, x_i)$$

calculated for the particle as though the oscillator were absent,

$$S_0 = \sum_{i=0}^{j-1} \left[ \frac{\epsilon}{2} \left( \frac{q_{i+1} - q_i}{\epsilon} \right)^2 - \frac{\epsilon \omega^2}{2} q_{i+1}^2 \right]$$

that of the oscillator alone, and

$$S_I = \sum_{i=0}^{j-1} \gamma_i q_i$$

(where  $\gamma_i = \gamma(x_i, t_i)$ ) is the action of interaction between the particle and the oscillator. The normalizing constant,  $a$ , for the oscillator is  $(2\pi\epsilon i/\hbar)^{-1/2}$ . Now the exponential depends quadratically upon all the  $q_i$ . Hence, the integrations over all the variables  $q_i$ , for  $0 < i < j$  can easily be performed. One is integrating a sequence of Gaussian integrals.

The result of these integrations is, writing  $T = t'' - t'$ ,  $(2\pi i \hbar \sin \omega T / \omega)^{-1/2} \exp(S_p + Q(q_i, q_0)) / \hbar$ , where  $Q(q_j, q_0)$  go turns out to be just the classical

---

<sup>24</sup>The generalization to the case that  $\gamma$  depends on the velocity,  $\dot{x}$ , of the particle presents no problem.

action for the forced harmonic oscillator (see reference 15). Explicitly it is

$$\begin{aligned} Q(q_j, q_0) = & \frac{\omega}{2 \sin \omega T} \left[ (\cos \omega T)(q_j^2 + q_0^2) - 2q_j q_0 \right. \\ & + \frac{2q_0}{\omega} \int_{t'}^{t''} \gamma(t) \sin \omega(t - t') dt + \frac{2q_j}{\omega} \int_{t'}^{t''} \gamma(t) \sin \omega(t'' - t) dt \\ & \left. - \frac{2}{\omega^2} \int_{t'}^{t''} \int_{t'}^t \gamma(t) \gamma(s) \sin \omega(t'' - t) \times \sin \omega(s - t') ds dt \right]. \end{aligned}$$

It has been written as though  $\gamma(t)$  were a continuous function of time. The integrals really should be split into Riemann sums and the quantity  $\gamma(x_i, t_i)$  substituted for  $\gamma(t_i)$ . Thus,  $Q$  depends on the coordinates of the particle at all times through the  $\gamma(x_i, t_i)$  and on that of the oscillator at times  $t'$  and  $t''$  only. Thus, the quantity (61) becomes

$$\begin{aligned} \langle \chi_{t''} \varphi_m | 1 | \psi_{t'} \varphi_n \rangle_{Sp+S_0+S_I} = & \int \dots \int \chi_{t''}^*(x_i) G_{mn} \times \\ & \times \exp \left( \frac{iS_p}{\hbar} \right) \psi_{t'}(x_0) \frac{dx_0}{A} \dots \frac{dx_{j-1}}{A} dx_i = \langle \chi_{t''} | G_{mn} | \psi_{t'} \rangle_{S_p} \end{aligned}$$

which now contains the coordinates of the particle only, the quantity  $G_{mn}$  being given by

$$G_{mn} = (2\pi i \hbar \sin \omega T / \omega)^{-1/2} \int \int \varphi_m^*(q_i) \times \exp(iQ(q_i, q_0)/\hbar) \varphi_n(q_0) dq_j dq_0.$$

Proceeding in an analogous manner one finds that all of the oscillators of the electromagnetic field can be eliminated from a description of the motion of the charges.

## Statistical Mechanics

### Spin and Relativity

Problems in the theory of measurement and statistical quantum mechanics are often simplified when set up from the point of view described here. For example, the influence of a perturbing measuring instrument can be integrated out in principle as we did in detail for the oscillator. The statistical density matrix has a fairly obvious and useful generalization. It results from considering the square of (38). It is an expression similar to (38) but containing integrations over two sets of variables  $dx_i$ , and  $dx'_i$ . The exponential

is replaced by  $\exp i(S - S')/\hbar$ , where  $S'$  is the same function of the  $x'_i$  as  $S$  is of  $x_i$ . It is required, for example, to describe the result of the elimination of the field oscillators where, say, the final state of the oscillators is unspecified and one desires only the sum over all final states  $m$ .

Spin may be included in a formal way. The Pauli spin equation can be obtained in this way: One replaces the vector potential interaction term in  $S(x_{i+1}, x_i)$ ,

$$\frac{e}{2c}(\mathbf{x}_{i+1} - \mathbf{x}_i) \cdot \mathbf{A}(\mathbf{x}_i) + \frac{e}{2c}(\mathbf{x}_{i+1} - \mathbf{x}_i) \cdot \mathbf{A}(\mathbf{x}_{i+1})$$

arising from expression (13) by the expression

$$\frac{e}{2c}(\sigma \cdot (\mathbf{x}_{i+1} - \mathbf{x}_i))(\sigma \cdot \mathbf{A}(\mathbf{x}_i)) + \frac{e}{2c}(\sigma \cdot \mathbf{A}(\mathbf{x}_{i+1}))(\sigma \cdot (\mathbf{x}_{i+1} - \mathbf{x}_i)).$$

Here  $\mathbf{A}$  is the vector potential,  $\mathbf{x}_{i+1}$  and  $\mathbf{x}$ , the vector positions of a particle at times  $t_{i+1}$  and  $t_i$ , and  $\sigma$  is Pauli's spin vector matrix. The quantity  $\Phi$  must now be expressed as  $\Pi_i \exp iS(x_{i+1}, x_i)/\hbar$  for this differs from the exponential of the sum of  $S(x_{i+1}, x_i)$ . Thus,  $\Phi$  is now a spin matrix.

The Klein Gordon relativistic equation can also be obtained formally by adding a fourth coordinate to specify a path. One considers a "path" as being specified by four functions  $x^{(\mu)}(\tau)$  of a parameter  $\tau$ . The parameter  $\tau$  now goes in steps  $\epsilon$  as the variable  $t$  went previously. The quantities  $x^{(1)}(t), x^{(2)}(t), x^{(3)}(t)$  are the space coordinates of a particle and  $x^{(4)}(t)$  is a corresponding time. The Lagrangian used is

$$\sum'_{\mu=1} [(dx^\mu/d\tau)^2 + (e/c)(dx^\mu/d\tau)\mathbf{A}_\mu],$$

where  $A_\mu$  is the 4-vector potential and the terms in the sum for  $\mu = 1, 2, 3$  are taken with reversed sign. If one seeks a wave function which depends upon  $\tau$  periodically, one can show this must satisfy the Klein Gordon equation. The Dirac equation results from a modification of the Lagrangian used for the Klein Gordon equation, which is analogous to the modification of the non-relativistic Lagrangian required for the Pauli equation. What results directly is the square of the usual Dirac operator.

These results for spin and relativity are purely formal and add nothing to the understanding of these equations. There are other ways of obtaining the Dirac equation which offer some promise of giving a clearer physical interpretation to that important and beautiful equation.

The author sincerely appreciates the helpful advice of Professor and Mrs. H. C. Corben and of Professor H. A. Bethe. He wishes to thank Professor J. A. Wheeler for very many discussions during the early stages of the work.

# A Generalized Theory of Gravitation

ALBERT EINSTEIN

*Institute for Advanced Study, Princeton, New Jersey*

**I**N the following we shall give a new presentation of the generalized theory of gravitation, which constitutes a certain progress in clarity as compared to the previous presentations.\* It is our aim to achieve a theory of the total field by a generalization of the concepts and methods of the relativistic theory of gravitation.

## 1. THE FIELD STRUCTURE

The theory of gravitation represents the field by a symmetric tensor  $g_{ik}$ , i.e.,  $g_{ik} = g_{ki}$  ( $i, k = 1, \dots, 4$ ), where the  $g_{ik}$  are real functions of  $x_1, \dots, x_4$ .

In the generalized theory the total field is represented by a Hermitian tensor. The symmetry property of the (complex)  $g_{ik}$  is

$$g_{ik} = \overline{g_{ki}}.$$

If we decompose  $g_{ik}$  into its real and imaginary components, then the former is a symmetric tensor ( $g_{ik}$ ), the latter an antisymmetric tensor ( $g_{ij}$ ). The  $g_{ik}$  are still functions of the real variables  $x_1, \dots, x_4$ .

The formally natural character of this generalization of the symmetric tensor becomes particularly noticeable by the following consideration: From the covariant vector  $A_i$  one can form through multiplication the particular symmetric covariant tensor  $A_i A_k$ . From such tensors every symmetric tensor of rank 2 can be obtained through summation with real coefficients:

$$g_{ik} = \sum_{\alpha} c_{\alpha} A_i A_k.$$

In an analogous manner we form from a complex vector  $A_i$  the special Hermitian tensor  $A_i \overline{A_k}$  (remains fixed if we interchange  $i$  and  $k$  and take the complex conjugate). We then get the representation of a general Hermitian tensor

\* A. Einstein, "A generalization of the relativistic theory of gravitation," Ann. Math. 46 (1945); A. Einstein and E. G. Straus, "A generalization of the relativistic theory of gravitation II," Ann. Math. 47 (1946).

of rank 2,

$$g_{ik} = \sum_{\alpha} c_{\alpha} A_i \overline{A_k},$$

where the  $c_{\alpha}$  are again real constants.

The determinant  $g = |g_{ik}| (\neq 0)$  is real.

Proof:

$$|g_{ik}| = |g_{ki}| = |\overline{g_{ik}}| = \overline{|g_{ik}|}.$$

We can associate a contravariant  $g^{ik}$  to the covariant  $g_{ik}$  just as in the case of real fields by setting

$$g_{is} g^{ls} = \delta_i^l \quad (\text{or } g_{si} g^{sl} = \delta_i^l),$$

where  $\delta_i^l$  is the Kronecker tensor. Here the order of indices is important and, for example,  $g_{is} g^{sl}$  does not equal  $\delta_i^l$ . In the following the tensor density  $g^{ik} = g^{ik}(g)^{\frac{1}{2}}$  plays an important role.

From a group theoretical point of view the introduction of a Hermitian tensor is somewhat arbitrary, since both individual additive components  $g_{ik}$  and  $g_{ij}$  have tensor character. However, this flaw is somewhat ameliorated by the fact that, just as in the case of real fields, there is a natural way of associating parallel translations to the Hermitian  $g_{ik}$ ; this is the main basis for the claim that the introduction of a Hermitian  $g_{ik}$  is natural.

## 2. INFINITESIMAL PARALLEL TRANSLATIONS, ABSOLUTE DIFFERENTIATION AND CURVATURE

In the theory of real fields we give the infinitesimal parallel translation of a vector  $A^i$  or  $A_i$  by

$$\left. \begin{aligned} \delta A^i &= -\Gamma^i_{st} A^s dx^t \\ \delta A_i &= \Gamma^s_{it} A_s dx_t \end{aligned} \right\} \quad (1)$$

with a corresponding introduction of infinitesimal parallel translations for tensors of higher rank.

The second equation of (1) is connected with the first by the demand that

$$0 = \delta(\delta^k_i) = (\delta^s_i \Gamma^k_{sl} - \delta^k_s \Gamma^s_{il}) dx^l.$$

From (1) we get in the well-known manner the tensor character of

$$dA^i - \delta A^i = \left( \frac{\partial A^i}{\partial x_t} + A^s \Gamma_{st}^i \right) dx^t,$$

which yields the concept of covariant differentiation

$$\begin{aligned} A_{i;t} &= \frac{\partial A^i}{\partial x_t} + A^s \Gamma_{st}^i \\ A_{i;t} &= \frac{\partial A_i}{\partial x_t} - A_s \Gamma_{it}^s \end{aligned} \quad (2)$$

In order to obtain the covariant derivative of  $g_{ik}$  we write

$$\begin{aligned} A_{i;l} &= \frac{\partial A_i}{\partial x_l} - A_s \Gamma_{il}^s, \\ A_{k;l} &= \frac{\partial A_k}{\partial x_l} - A_s \Gamma_{kl}^s, \end{aligned}$$

multiplying the first equation by  $A_k$ , the second by  $A_i$  and adding we get

$$\begin{aligned} A_i A_{k;l} + A_k A_{i;l} &= (A_i A_k)_l \\ &= (A_i A_k)_l - (A_s A_k) \Gamma_{il}^s - (A_i A_s) \Gamma_{kl}^s, \end{aligned}$$

and since  $g_{ik}$  can be constructed as the sum of such special tensors we get

$$g_{ik;l} = g_{ik,l} - g_{sk} \Gamma_{il}^s - g_{is} \Gamma_{kl}^s.$$

The  $\Gamma$  are now determined from the  $g$  and their first derivatives by the demand that the absolute derivative of the  $g_{ik}$  vanish

$$0 = g_{ik;l} - g_{sk} \Gamma_{il}^s - g_{is} \Gamma_{kl}^s. \quad (3)$$

However, since the  $g_{ik}$  are symmetric, these are only 40 equations for the 64  $\Gamma$ . In order to complete the determination of the  $\Gamma$  one uses the only possible invariant algebraic condition, namely, the condition of symmetry

$$\Gamma_{ik}^l = \Gamma_{ki}^l. \quad (4)$$

We now transfer this development to the complex case by defining parallel translation as in (1). However, this gives rise to a certain complication, since if we start from the translation of a complex vector,

$$\delta A^i = \Gamma_{it}^s A_s dx^t,$$

where the  $\Gamma$  will, in general, also be complex, and pass to the complex conjugate of this equation

$$\overline{\delta A_i} = \overline{\Gamma_{it}^s A_s} dx^t,$$

then we see that we have there an equation which also defines a parallel translation, but this parallel translation may differ from the first. We define then two kinds of parallel translation

$$\begin{aligned} \delta A_+^i &= -\Gamma_{st}^i A^s dx^t \\ \delta A_-^i &= \Gamma_{st}^i A_s dx^t \end{aligned} \quad (1a)$$

and

$$\begin{aligned} \delta A_-^i &= -\overline{\Gamma_{st}^i A^s} dx^t \\ \delta A_+^i &= \overline{\Gamma_{st}^i A_s} dx^t \end{aligned} \quad (1b)$$

and, correspondingly, two kinds of covariant differentiation  $A_+^i$ ,  $A_{-i}$ , and  $A_+^i$ ,  $A_{-i}$  as in (2).

From (1a) and (1b) we get

$$\delta \overline{A_-^i} = \overline{\delta A_+^i} \quad \text{and} \quad \delta \overline{A_+^i} = \overline{\delta A_-^i}$$

*In order that conjugate vectors have conjugate translations and derivatives it is necessary upon passage to the conjugate to change the character of translation or of differentiation, i.e., to pass to the conjugate  $\Gamma$ .* In order to obtain the covariant derivative of a Hermitian tensor we write in analogy to the real case:

$$\begin{aligned} A_{+i;l} &= \frac{\partial A_i}{\partial x_l} - A_s \Gamma_{il}^s, \\ \overline{A_{-k;l}} &= \frac{\partial \overline{A_k}}{\partial x_l} - \overline{A_s \Gamma_{kl}^s}. \end{aligned}$$

From this we get as before

$$\begin{aligned} A_i \overline{A_{-k;l}} + \overline{A_k} A_{+i;l} &= (A_i A_k)_l \\ &= (A_i A_k)_l - (A_s \overline{A_k}) \Gamma_{il}^s - (A_i \overline{A_s}) \overline{\Gamma_{kl}^s}, \end{aligned}$$

and since  $g_{ik}$  can be constructed as the sum of such special tensors we get

$$g_{+i;k;l} = g_{ik;l} - g_{sk} \Gamma_{il}^s - g_{is} \overline{\Gamma_{kl}^s}.$$

The analog to (3) is the requirement that this absolute derivative vanish

$$0 = g_{+i;k;l} = g_{ik;l} - g_{sk} \Gamma_{il}^s - g_{is} \overline{\Gamma_{kl}^s}. \quad (3a)$$

These equations are Hermitian in the indices  $i, k$  (go into themselves if we interchange  $i, k$  and pass to the conjugate complex) and therefore again do not suffice to determine the complex  $\Gamma$ . In analogy to (4) we have as the only possible invariant algebraic determination the condition of Hermiticity

$$\Gamma^l_{ik} = \overline{\Gamma^l_{ki}}. \quad (4a)$$

Instead of (3a) we can then write

$$0 = g_{+k;l} = g_{ik,l} - g_{sk}\Gamma^s_{il} - g_{is}\Gamma^s_{lk}, \quad (3b)$$

which implies both (3a) and (4a).

*Absolute differentiation of vector densities.* If we multiply (3b) by  $\frac{1}{2}g^{ik}$  and sum over  $i$  and  $k$ , then we get the vector equation

$$\frac{1}{(g)^{\frac{1}{2}}} \frac{\partial(g)^{\frac{1}{2}}}{\partial x_l} - \frac{1}{2}(\Gamma^a_{al} + \Gamma^a_{la}) = 0,$$

or shorter

$$\frac{\partial(g)^{\frac{1}{2}}}{\partial x_l} - (g)^{\frac{1}{2}}\Gamma^a_{la} = 0. \quad (3c)$$

$(g)^{\frac{1}{2}}$  is a scalar density, the left side of (3c) is a vector density. The latter will also hold if  $(g)^{\frac{1}{2}}$  is replaced by an arbitrary scalar density  $\rho$ . We may therefore introduce as the absolute derivative of a scalar density  $\rho$ :

$$\rho_{;l} = \rho_{,l} - \rho\Gamma^a_{la}. \quad (5)$$

This permits us to introduce absolute differentiation of tensor densities.

Example: If we multiply the right side of the equation

$$A^i_{+;l} = A^i_{,l} + A^s\Gamma^i_{sl}$$

by a scalar density  $\rho$ , then we get the tensor density

$$(\rho A^i)_{,l} + (\rho A^s)\Gamma^i_{sl} - A^i\rho_{,l}$$

or, after introducing the vector density  $\mathfrak{A}^i = \rho A^i$

$$\mathfrak{A}^i_{+;l} = \mathfrak{A}^i_{,l} + \mathfrak{A}^s\Gamma^i_{sl} - \mathfrak{A}^i \frac{\rho_{,l}}{\rho},$$

or according to (5)

$$(\mathfrak{A}^i_{+;l} + \mathfrak{A}^s\Gamma^i_{sl} - \mathfrak{A}^i\Gamma^a_{la}) - \mathfrak{A}^i\rho_{,l}.$$

Since the last term is a tensor density, the term in brackets is also a tensor density which we may

define as the absolute derivative  $\mathfrak{A}^i_{+;l}$  of a vector density  $\mathfrak{A}^i$ :

$$\mathfrak{A}^i_{+;l} = \mathfrak{A}^i_{,l} + \mathfrak{A}^s\Gamma^i_{sl} - \mathfrak{A}^i\Gamma^a_{la}. \quad (6)$$

In an analogous manner we may define the absolute derivatives of arbitrary tensor densities. They differ from the absolute derivative of the tensor by a last term like  $-\mathfrak{A}^i\Gamma^a_{la}$ .

Just as in the case of real fields we can bring (3a) into a contravariant form; however, we have to be careful about the order of indices. We obtain the equivalent equations

$$0 = g^i_k \mathfrak{A}_{+;l} = g^{ik}_{,l} + g^{sk}\Gamma^i_{sl} + g^{is}\Gamma^k_{ls}, \quad (3d)$$

or, after introducing the contravariant tensor density,  $\mathfrak{g}^{ik} = g^{ik}(g)^{\frac{1}{2}}$

$$0 = g^i_k \mathfrak{A}_{+;l} = g^{ik}_{,l} + g^{sk}\Gamma^i_{sl} + g^{is}\Gamma^k_{ls} - g^{ik}\Gamma^s_{ls}. \quad (3e)$$

The Eqs. (3a), (3d), and (3e) are equivalent.

*Curvature:* The change which a vector undergoes upon parallel translation around the boundary curve of an infinitesimal element of area has vector character. This leads to the formation of a curvature tensor also in the case of our generalized field. We have here the choice whether to use a “+” translation or a “-” translation; however, the results of the two translations are conjugate complex, so that it suffices to consider one form.

We obtain the tensor

$$R^i_{klm} = \Gamma^i_{kl,m} - \Gamma^i_{km,l} - \Gamma^i_{al}\Gamma^a_{km} + \Gamma^i_{am}\Gamma^a_{kl}, \quad (7)$$

and the corresponding contracted tensor (contraction with respect to  $i$  and  $m$ )

$$R^*_{kl} = \Gamma^a_{kl,a} - \Gamma^a_{ka,l} - \Gamma^a_{kb}\Gamma^b_{al} + \Gamma^a_{kl}\Gamma^b_{ab}. \quad (8)$$

There also exists a non-vanishing contraction with respect to  $i$  and  $k$  which yields the tensor

$$\Gamma^a_{al,m} - \Gamma^a_{am,l}. \quad (9)$$

However, we shall not use this tensor as we shall justify later. The tensor  $R^*_{kl}$  is not Hermitian. We form the Hermitian tensor  $R_{ik} = \frac{1}{2}(R^*_{ik} + R^*_{ki})$ . We thus get

$$R_{ik} = \Gamma^a_{ik,a} - \frac{1}{2}(\Gamma^a_{ia,k} + \Gamma^a_{ak,i}) - \Gamma^a_{ib}\Gamma^b_{ak} + \Gamma^a_{ik}\Gamma^b_{ab}. \quad (8a)$$

### 3. HAMILTONIAN PRINCIPLE. FIELD EQUATIONS

In the case of the real symmetric field one obtains the field equations most simply in the following manner. We use as Hamilton function the scalar density

$$\mathfrak{H} = g^{ik} R_{ik}. \quad (10)$$

If we vary the volume integral of  $\mathfrak{H}$  independently with respect to  $\Gamma$  and  $g$ , then (in the case of real fields) variation with respect to  $\Gamma$  yields Eq. (3), and variation with respect to  $g$  yields the equations  $R_{ik} = 0$ . If we apply the same method to our case of a complex field (where  $\mathfrak{H}$  is still real) then we see a complication, since the variation with respect to  $\Gamma$  does not immediately yield Eq. (3a), which we wish to keep in any case. The variation with respect to  $\Gamma$  yields

$$\begin{aligned} -\{g^{ik}_{,a} + g^{sk}\Gamma^i_{sa} + g^{is}\Gamma^k_{as} - g^{ik}\Gamma^b_{ab}\} \\ + \frac{1}{2}\{g^{is}_{,s} + g^{st}\Gamma^i_{st} - g^{is}\Gamma^a_{sa}\}\delta_a^k \\ + \frac{1}{2}\{g^{sk}_{,s} + g^{st}\Gamma^k_{st} + g^{sk}\Gamma^a_{sa}\}\delta_a^i \\ + \frac{1}{2}\{g^{is}\Gamma^a_{sa}\delta_a^k - g^{sk}\Gamma^a_{sa}\delta_a^i\}. \end{aligned} \quad (11)$$

The first bracket is  $g^{ik}_{,a}$ ; the second and third brackets are contractions of this quantity. If there were no fourth bracket then (11) would imply the vanishing of  $g^{ik}_{,a}$ , that is, (3a). However, this would require the vanishing of  $\Gamma^a_{sa}$  to which demand we have no right for the time being.

We can resolve this difficulty in the following manner. We form the imaginary part of (11):

$$\begin{aligned} -g^{ik}_{,a} - g^{sk}\Gamma^i_{sa} - g^{is}\Gamma^k_{sa} \\ - g^{is}\Gamma^k_{as} - g^{sk}\Gamma^k_{as} + g^{ik}\Gamma^b_{ab} \\ + \frac{1}{2}g^{is}_{,s}\delta_a^k + \frac{1}{2}g^{sk}_{,s}\delta_a^i = 0. \end{aligned}$$

If we contract this equation with respect to  $k$  and  $a$  we get

$$\frac{1}{2}g^{is}_{,s} + g^{is}\Gamma^a_{sa} = 0. \quad (11a)$$

From this we can deduce that the necessary and sufficient\*\* condition for the vanishing of the  $\Gamma^a_{sa}$  is the vanishing of the  $g^{is}_{,s}$ . In order to

\*\* This holds for all points if we demand that the  $\Gamma$  be continuous and determined uniquely by the equations (3b); because then the determinant  $[g^{is}]$  can vanish nowhere.

satisfy this *identically* it suffices to assume

$$g^{is} = g^{ist},_t, \quad (12)$$

where  $g^{ist}$  is a tensor density which is antisymmetric in all three indices. That is, we require that  $g^{is}$  be derived from a "vector potential." We therefore substitute in the Hamilton function

$$g^{ik} = g^{ik} + g^{ikl},_l \quad (13)$$

and vary independently with respect to the  $\Gamma$ ,  $g^{ik}$  and  $g^{ikl}$ . The variation with respect to the  $\Gamma$  then yields (3a), as we have shown. The variation with respect to the other quantities yields the equations

$$R_{ik} = 0, \quad (14)$$

$$R_{ik,l} + R_{kl,i} + R_{li,k} = 0. \quad (15)$$

In addition, we have the equations

$$g^{ik}_{,l} = 0 \quad \text{or} \quad g_{ik;l} = 0, \quad (3a)$$

$$\Gamma^s_{is} = 0, \quad (16)$$

$$g^{is}_{,s} = 0 \quad \text{or} \quad g^{is} = g^{ist},_t. \quad (17)$$

Considering (3a), each of the systems (16), (17) implies the other; this is proven by showing that (3a) implies the equation

$$g^{is}_{,s} - g^{is}\Gamma^t_{st} = 0.$$

The system of field equations is therefore not weakened if we omit (17).

This is worth mentioning also for the following reason. While in the given derivation of the equations, special emphasis is given to the density  $g^{ik}$  rather than to the tensor  $g_{ik}$  (or  $g^{ik}$ ), the resulting system itself is free of such discrimination.

We now see that because of (16) the tensor (9) reduces to  $\Gamma^a_{al,m} - \Gamma^a_{am,l}$ , which vanishes because of Eq. (3c).

The derivation used here has the advantage, as compared to the previous one, that the Hamiltonian principle used is one without side conditions. This behavior is the same as that encountered in a (specially relativistic) derivation of Maxwell's equations from a variational principle. There (for imaginary time coordinate) the Hamiltonian function is  $\mathfrak{H} = \varphi_{\psi}\varphi_{\psi}$ . If we set here  $\varphi_{\psi} = \varphi_{i,k} - \varphi_{k,i}$  and vary with respect

to the  $\varphi_i$ , then we get the one system of equations ( $\varphi_{ik,k}=0$ ) directly, the other through elimination of the  $\varphi_i$ . This method corresponds to the one used above. One may, however, avoid the introduction of the potentials  $\varphi_i$  and instead adjoin the system of equations

$$\varphi_{ik,i} + \varphi_{kl,i} + \varphi_{lj,k} = 0$$

as side conditions for the  $\varphi_k$  in the variation. This corresponds to the treatment of  $g^{ik}_{\psi,\epsilon}=0$  as side condition for the variation in the previous paper. The side condition  $\Gamma^s_{\psi}=0$  which was introduced there could have been omitted.

#### REMARKS

In order to preserve the special character of locally space-like and time-like directions it is essential that the index of inertia of  $g_{ik}dx^idx^k$  be the same everywhere, i.e., that the determinant  $|g_{ik}|$  vanish nowhere. This can indeed be deduced from the requirement that the  $\Gamma$ -field be finite and determined everywhere by Eq. (3a). My assistant has given the following simple proof of this:

If the determinant  $|g_{ik}|$  should vanish in a point  $P$  then there would exist a vector  $\xi^s$  different from 0, such that  $g_{is}\xi^s=0$ . We now consider the real part of Eq. (3a):

$$g_{ik,i} - g_{sk}\Gamma^s_{il} - g_{is}\Gamma^s_{lk} - g_{sk}\Gamma^s_{il} - g_{is}\Gamma^s_{lk} = 0.$$

If we multiply this equation (at the point  $P$ ) by  $\xi^i\xi^k\xi^l$  and sum over  $i, k, l$ , then the second and third terms vanish by definition of  $\xi$ , and the fourth and fifth because of the antisymmetry of the  $\Gamma$ . There exists, therefore, a linear combination of Eq. (3a) which does not contain the  $\Gamma$ . Hence at such a point the  $\Gamma$  either become infinite or not completely determined, in contradiction to our requirement.

Concerning the physical interpretation we re-

mark that the antisymmetric density  $g^{ikl}$  plays the role of an electromagnetic vector potential, the tensor  $g_{\psi,i} + g_{kj,i} + g_{lj,k}$  the role of current density. The latter quantity is the "complement" of a contravariant vector density with (identically) vanishing divergence.

Above we have used complex fields. However, there exists a theoretical possibility in which the  $g_{ik}$  and  $\Gamma^l_{ik}$  are real though not symmetric. Thus one can obtain a theory which in its final formulas corresponds, except for certain signs, to the one developed above. E. Schrödinger, too, has based his affine theory (i.e., based on the  $\Gamma$  as fundamental field quantities) on real fields. I therefore wish to give here some formal reasons for the preferability of complex fields.

A Hermitian tensor  $g_{ik}$  can be constructed additively from vectors according to the scheme  $g_{ik} = \sum_{\alpha} c A_i \overline{A_k}$ . The essential fact here is that with the use of *one* complex vector  $A_i$  one can construct the Hermitian tensor  $A_i A_k$  through multiplication, which is a close analogy to the case of symmetric real fields. A non-symmetric real tensor cannot be constructed from vectors in such close analogy.

We now consider translation quantities  $\Gamma^l_{ik}$  which are not symmetric in the lower indices. To them we have in both the real and the complex cases the adjoined ("conjugate") translation quantities  $\tilde{\Gamma}^l_{ik} = \Gamma^l_{ki}$ . In the complex case we have associated with the parallel translation of a vector

$$\delta A^i = -\Gamma^i_{st} A^s dx^t$$

the parallel translation of its conjugate complex vector

$$\delta \overline{A}^i = -\overline{\Gamma^i_{st}} \overline{A}^s dx^t.$$

Hence in the case of complex fields the adjoined translation corresponds to adjoined objects, while in the case of real fields there is no such adjoined object.



## Letters to the Editor

**P**UBLICATION of brief reports of important discoveries in physics may be secured by addressing them to this department. The closing date for this department is five weeks prior to the date of issue. No proof will be sent to the authors. The Board of Editors does not hold itself responsible for the opinions expressed by the correspondents. Communications should not exceed 600 words in length.

### The Origin of Chemical Elements

R. A. ALPHER\*

Applied Physics Laboratory, The Johns Hopkins University,  
Silver Spring, Maryland

AND

H. BETHE

Cornell University, Ithaca, New York

AND

G. GAMOW

The George Washington University, Washington, D. C.

February 18, 1948

**A**S pointed out by one of us,<sup>1</sup> various nuclear species must have originated not as the result of an equilibrium corresponding to a certain temperature and density, but rather as a consequence of a continuous building-up process arrested by a rapid expansion and cooling of the primordial matter. According to this picture, we must imagine the early stage of matter as a highly compressed neutron gas (overheated neutral nuclear fluid) which started decaying into protons and electrons when the gas pressure fell down as the result of universal expansion. The radiative capture of the still remaining neutrons by the newly formed protons must have led first to the formation of deuterium nuclei, and the subsequent neutron captures resulted in the building up of heavier and heavier nuclei. It must be remembered that, due to the comparatively short time allowed for this process,<sup>1</sup> the building up of heavier nuclei must have proceeded just above the upper fringe of the stable elements (short-lived Fermi elements), and the present frequency distribution of various atomic species was attained only somewhat later as the result of adjustment of their electric charges by  $\beta$ -decay.

Thus the observed slope of the abundance curve must not be related to the temperature of the original neutron gas, but rather to the time period permitted by the expansion process. Also, the individual abundances of various nuclear species must depend not so much on their intrinsic stabilities (mass defects) as on the values of their neutron capture cross sections. The equations governing such a building-up process apparently can be written in the form:

$$\frac{dn_i}{dt} = f(t)(\sigma_{i-1}n_{i-1} - \sigma_i n_i) \quad i=1, 2, \dots, 238, \quad (1)$$

where  $n_i$  and  $\sigma_i$  are the relative numbers and capture cross sections for the nuclei of atomic weight  $i$ , and where  $f(t)$  is a factor characterizing the decrease of the density with time.

We may remark at first that the building-up process was apparently completed when the temperature of the neutron gas was still rather high, since otherwise the observed abundances would have been strongly affected by the resonances in the region of the slow neutrons. According to Hughes,<sup>2</sup> the neutron capture cross sections of various elements (for neutron energies of about 1 Mev) increase exponentially with atomic number halfway up the periodic system, remaining approximately constant for heavier elements.

Using these cross sections, one finds by integrating Eqs. (1) as shown in Fig. 1 that the relative abundances of various nuclear species decrease rapidly for the lighter elements and remain approximately constant for the elements heavier than silver. In order to fit the calculated curve with the observed abundances<sup>3</sup> it is necessary to assume the integral of  $\rho_n dt$  during the building-up period is equal to  $5 \times 10^4$  g sec./cm<sup>3</sup>.

On the other hand, according to the relativistic theory of the expanding universe<sup>4</sup> the density dependence on time is given by  $\rho \leq 10^6/t^2$ . Since the integral of this expression diverges at  $t=0$ , it is necessary to assume that the building-up process began at a certain time  $t_0$ , satisfying the relation:

$$\int_{t_0}^{\infty} (10^6/t^2) dt \leq 5 \times 10^4, \quad (2)$$

which gives us  $t_0 \leq 20$  sec. and  $\rho_0 \leq 2.5 \times 10^5$  g sec./cm<sup>3</sup>. This result may have two meanings: (a) for the higher densities existing prior to that time the temperature of the neutron gas was so high that no aggregation was taking place, (b) the density of the universe never exceeded the value  $2.5 \times 10^5$  g sec./cm<sup>3</sup> which can possibly be understood if we

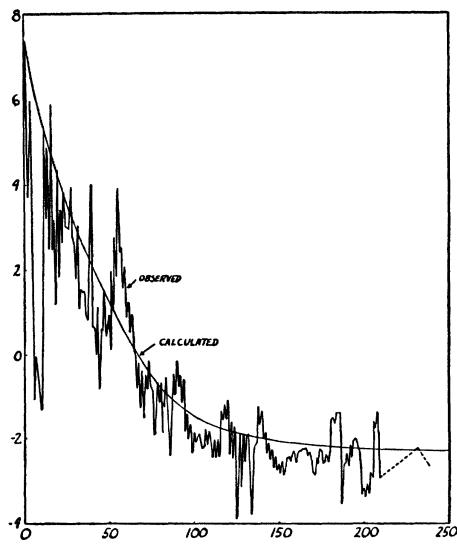


FIG. 1.  
Log of relative abundance  
Atomic weight

use the new type of cosmological solutions involving the angular momentum of the expanding universe (spinning universe).<sup>5</sup>

More detailed studies of Eqs. (1) leading to the observed abundance curve and discussion of further consequences will be published by one of us (R. A. Alpher) in due course.

\* A portion of the work described in this paper has been supported by the Bureau of Ordnance U. S. Navy, under Contract NOrd-7386.

<sup>1</sup> G. Gamow, Phys. Rev. **70**, 572 (1946).

<sup>2</sup> D. J. Hughes, Phys. Rev. **70**, 106(A) (1946).

<sup>3</sup> V. M. Goldschmidt, *Geochemische Verleitungsgegesetz der Elemente und der Atom-Arten*. IX. (Oslo, Norway, 1938).

<sup>4</sup> See, for example: R. C. Tolman, *Relativity, Thermodynamics and Cosmology* (Clarendon Press, Oxford, England, 1934).

<sup>5</sup> G. Gamow, Nature, October 19 (1946).

### A Beta-Ray Spectrometer Design of Quadratic Resolution-Solid Angle Relationship

S. FRANKEL

Frankel and Nelson, Los Angeles, California

February 16, 1948

In a  $\beta$ -spectrometer for use with low intensity sources it is advantageous to collect electrons emitted by the source in as large as possible a solid angle consistent with the required resolution. In conventional spectrometers the usable solid angle,  $\Omega$ , is proportional to the momentum spread,  $\delta p/p$ , for small  $\delta p$ . ( $\delta p$  is the half-intensity width observed for a point source of monoenergetic electrons.) The double focusing spectrometer<sup>1</sup> has a more favorable proportionality constant than the constant field magnetic lens ("solenoid") spectrometer. The thin-lens spectrometer has a still less favorable constant.<sup>2</sup> Figure 1 shows approximate  $\Omega$  vs.  $\delta p/p$  curves for these designs.

Witcher<sup>3</sup> has shown that the solenoid spectrometer brings monoenergetic rays having nearly the same initial angles with the axis,  $\gamma$ , to a "ring-focus" between the source and counter, nearer the latter (Fig. 2). By placing baffles inside and outside this ring-focus the resolution may be improved without decreasing  $\Omega$ . The resolution attainable is approximately that shown in Fig. 1 for rays with  $30^\circ < \gamma < 60^\circ$ , somewhat poorer outside this range. For

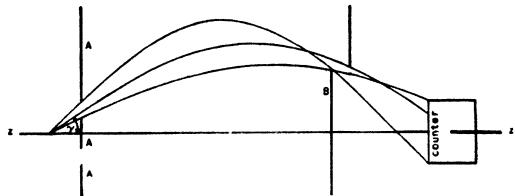


FIG. 2. Paths of electrons in a homogeneous magnetic field.  $z-z$ , axis of symmetry. Azimuthal motion of electrons not indicated. (A) Baffles defining range of  $\gamma$ . (B) Ring-focus baffles.

small  $\Omega$ ,  $\delta p = 0(\Omega^2)$ . Since the improvement in resolution attainable in this way seems not to be widely appreciated, it may be useful to direct attention to it.

Changing the energy of the electrons (or the field strength) without change in the range of  $\gamma$  uniformly expands or contracts the paths shown in Fig. 2 about the source as the fixed point. The best resolution is therefore obtained by placing the ring-focus baffles so that their defining edges lie on a cone with vertex at the source and axis parallel to the magnetic field.

It seems likely that a similar ring-focus exists in a thin-lens spectrometer and has similar favorable properties. Thus it is probably possible to combine the copper and power efficiency of the thin-lens design with a favorable resolution vs. solid angle curve. The position and properties of this ring-focus could be found experimentally (e.g., by the use of moveable baffles) or by numerical integration of the electron path equations.

The source diameter just sufficient to impair the momentum resolution is of the order of  $(\delta p/p) \cdot \tan \gamma \cdot (\text{radius of curvature})$  for the solenoid spectrometer either with or without the ring-focus baffles. Thus when an extended source is desirable (e.g., with a source of low specific activity) the improvement in counting rate at fixed resolution shown in Fig. 2 is genuine, while the improvement in resolution at fixed counting rate is in part specious.

<sup>1</sup> Siegbahn and Svartholm, Nature **157**, 872 (1946).

<sup>2</sup> T. Lauritsen, private communication.

<sup>3</sup> Clifford M. Witcher, Phys. Rev. **60**, 32 (1941).

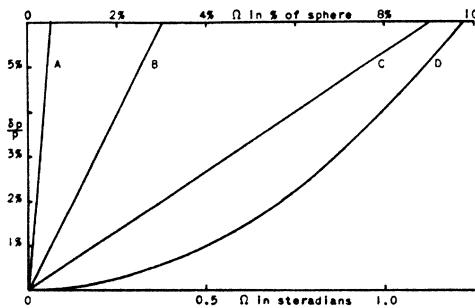


FIG. 1. Momentum resolution,  $\delta p/p$ , vs. solid angle,  $\Omega$ , of rays used: (A) A typical thin-lens spectrometer,<sup>2</sup> (B) Solenoid spectrometer for small  $\gamma$ , (C) The Siegbahn-Svartholm double focusing spectrometer. (D) The ring-focus baffled solenoid spectrometer. All curves are approximate and refer only to a point source.

## Relativistic Cut-Off for Quantum Electrodynamics

RICHARD P. FEYNMAN  
*Cornell University, Ithaca, New York*  
 (Received July 12, 1948)

A relativistic cut-off of high frequency quanta, similar to that suggested by Bopp, is shown to produce a finite invariant self-energy for a free electron. The electromagnetic line shift for a bound electron comes out as given by Bethe and Weisskopf's wave packet prescription. The scattering of an electron in a potential, without radiation, is discussed. The cross section remains finite. The problem of polarization of the vacuum is not solved. Otherwise, the results will in general agree essentially with those calculated by the prescription of Schwinger. An alternative cut-off procedure analogous to one proposed by Wataghin, which eliminates high frequency intermediate states, is shown to do the same things but to offer to solve vacuum polarization problems as well.

THE main problems of quantum electrodynamics have been essentially solved by the observations of Bethe<sup>1</sup> and of Weisskopf<sup>2</sup> that the divergent terms in the line shift problem can be thought to be contained in a renormalization of the mass of a free electron. That this principle applies as well to other problems was demonstrated by Lewis<sup>3</sup> in analyzing the radiationless scattering of an electron in a potential. Ambiguities which remained in the subtraction procedures are removed by Schwinger.<sup>2,4</sup> He formulated, in a general way, which terms are to be identified in a future correct theory with rest mass, and hence should be omitted from a calculation which does not renormalize the mass. These results are remarkable because they solve the problem without the addition of any new fundamental lengths or dimensions.

The solution given by Schwinger does, however, assume that in some future theory the divergent self-energy terms will be finite. Therefore, it is of interest to point out that there is a model, a modification of ordinary electrodynamics, for which all quantities automatically do come out finite. With this model the ideas of Bethe, Oppenheimer, and Lewis and Schwinger can be directly confirmed.

The model results from the quantization of a classical theory described in a previous paper.<sup>5</sup>

In this paper we describe only the results for processes in which only virtual quanta are emitted and absorbed. The problems of permanent emission and the position of positron theory must be more completely studied. It is hoped that a complete physical theory may be published in the near future. Lacking such a complete picture, the present paper may be looked upon merely as presenting an arbitrary rule to cut off at high frequencies in a relativistically invariant manner, the otherwise divergent integrals appearing in quantum field theories. For electrodynamics the rule is to consider the (positive) frequency  $\omega$  and wave number  $\mathbf{k}$  of the field oscillators as independent and to integrate them over the density function  $g(\omega^2 - k^2)d\omega dk$  where

$$g(\omega^2 - k^2) = \int_0^\infty [\delta(\omega^2 - k^2) - \delta(\omega^2 - k^2 - \lambda^2)]G(\lambda)d\lambda. \quad (1)$$

Here  $\delta(x)$  is Dirac's delta function and  $G(\lambda)$  is some smooth function such that  $\int_0^\infty G(\lambda)d\lambda = 1$  and for which the mean values of  $\lambda$  which are important are of order of the frequency 137  $mc^2/\hbar$ , or higher. Ordinary quantum electrodynamics replaces the function  $g(\omega^2 - k^2)$  by  $\delta(\omega^2 - k^2)$ . According to (1), the density  $g$  is not everywhere positive.<sup>5</sup> Therefore, the model is essentially that due to Bopp.<sup>6</sup>

The model therefore contains an arbitrary function and the numerical results depend on the

<sup>1</sup>H. A. Bethe, Phys. Rev. **72**, 339 (1947); **73**, 1271A (1948).

<sup>2</sup>J. Schwinger and V. Weisskopf, Phys. Rev. **73**, 1272A (1948).

<sup>3</sup>H. W. Lewis, Phys. Rev. **73**, 173 (1948).

<sup>4</sup>J. Schwinger, Phys. Rev. **73**, 415A (1948).

<sup>5</sup>R. P. Feynman, Phys. Rev. **74**, 939 (1948).

<sup>6</sup>F. Bopp, Ann. d. Physik **42**, 573 (1942).

form of  $G(\lambda)$ . However, the only term that depends seriously (logarithmically) on the cut-off frequency is the self-energy, which can be used to renormalize the electron mass. After this is done, the remaining terms are nearly independent of the function  $G(\lambda)$ .

We shall illustrate these points by studying the particular examples of self-energy and radiationless scattering. We shall then discuss an alternative cut-off procedure in which the density of electron states is cut off rather than that of the quanta. This promises to solve problems of vacuum polarization which are not touched by the former procedure.

#### SELF-ENERGY

The transverse self-energy of a free electron, of mechanical mass  $\mu$ , in state of momentum  $\mathbf{P}_0$  energy  $E_0 = (\mu^2 + P_0^2)^{\frac{1}{2}}$  is given to the first order in  $e^2$  by the second-order perturbation theory, using the one-electron theory of Dirac, by

$$\Delta E = -\frac{e^2}{4\pi^2} \sum_i \int \frac{d\mathbf{k}}{k} \left\{ \sum_+ \frac{(0|\alpha_i|f)(f|\alpha_i|0)}{E_f - E_0 + k} + \sum_- \frac{(0|\alpha_i|f)(f|\alpha_i|0)}{-E_f - E_0 + k} \right\}. \quad (2)$$

Here the intermediate state  $f$  arises from the initial state through emission of a quantum of momentum  $\mathbf{k}$  and of energy  $k = |\mathbf{k}|$  (the velocity of light is taken as unity, as is Planck's constant). Thus in the intermediate state the electron has momentum  $\mathbf{P}_f = \mathbf{P}_0 - \mathbf{k}$  and an energy of magnitude  $E_f = +(\mu^2 + P_f^2)^{\frac{1}{2}}$  but which may be either plus or minus in sign. The sums indicate the sum over all such intermediate states (actually just two) for each sign of the energy. The terms for positive and negative energy have been separated and the sums are written  $\sum_+$  and  $\sum_-$  for these two cases. The  $(f|\alpha_i|0)$  are the matrix elements of Dirac's  $\alpha$ -matrices, the sum on  $i$  being over the two directions of polarization of the quanta. We shall henceforth write the integral  $d\mathbf{k}/k$  over  $\mathbf{k}$  space by its equivalent  $2 \int d\omega d\mathbf{k} \delta(\omega^2 - k^2)$ , the integral being over all *positive*  $\omega$ , and all wave numbers  $\mathbf{k}$ . We shall also write  $\omega$  for  $k$  in the energy denominators as we shall later wish to distinguish the energy of a quantum and the magnitude of the momentum change that its

recoil represents. We may further simplify the expression by the use of the well-known projection operators:

$$\Lambda_f^\pm = (E_f \pm H_f)/2E_f = (E_f \pm \boldsymbol{\alpha} \cdot \mathbf{P}_f \pm \beta\mu)/2E_f.$$

According to the theory of holes, the last term, the transition to negative energy states, is to be left out; such transitions are prevented because the negative levels are already occupied. On the other hand, in the vacuum, electrons in state of energy  $-E_f$  could make virtual transitions to positive energy state  $E_0$ . This is now prevented by the presence of an electron in the state  $E_0$ , so that, relative to the vacuum, the transverse self-energy is

$$\Delta E = -\frac{e^2}{2\pi^2} \sum_i \int d\omega d\mathbf{k} \delta(\omega^2 - k^2) \times \left\{ \frac{(0|\alpha_i \Lambda_f^+ \alpha_i|0)}{E_f - E_0 + \omega} - \frac{(0|\alpha_i \Lambda_f^- \alpha_i|0)}{E_f + E_0 + \omega} \right\}. \quad (3)$$

The treatment of the longitudinal self-energy is usually different, for the longitudinal oscillators are first eliminated from the Hamiltonian, their effect being the term  $e^2/r_{00}$  where  $r_{00}$  is the meaningless distance of the electron from itself. These terms must be expressed as integrals over oscillators and combined with (3) before the change suggested by (1) is to be performed. An additional point of confusion is that the longitudinal elimination assumes the intermediate states to form a complete set as they do in (2), but the situation in (3) is not so clear. Fortunately, all these points may be most easily circumvented by simply not eliminating the longitudinal oscillators from the field Hamiltonian at all. One need simply to specify that the sum on  $i$  in (3) now be interpreted to mean the sum over each of three perpendicular space directions minus a term for the time direction. We may write  $\sum_i \alpha_i \Lambda \alpha_i = \boldsymbol{\alpha} \cdot \boldsymbol{\Lambda} \boldsymbol{\alpha} - \Lambda$ , which is a relativistic combination since  $\alpha_4 = 1$ . One does not need to be concerned about the gauge condition in a problem in which all quanta are virtual, for the quanta are created by a charge which is conserved. This solution automatically insures the gauge condition just as the Lienard Wiechert classical solution of the Maxwell equations will automatically satisfy the gauge

condition if the charge which produces the potential is conserved.

With this convention for  $\sum_i$ , Eq. (3) represents the total self-energy. It is easily calculated. The numerator of first term may be written as  $1/2E_f$  times  $\sum_i(0|\alpha_i(H_f+E_f)\alpha_i|0)$  where  $H_f$  is  $\alpha \cdot \mathbf{P}_f + \beta\mu$ . Now since  $\sum_i \alpha_i \alpha_i = +2$ ,  $\sum_i \alpha_i \beta \alpha_i = -4\beta$ , and  $\sum_i \alpha_i \alpha_i \alpha_i = -2\alpha$ , this becomes

$$-2(0| -E_f + 2\beta\mu + \alpha \cdot \mathbf{P}_f |0).$$

The diagonal elements of  $\beta$  and  $\alpha$  for the state 0 are  $\mu/E_0$  and  $\mathbf{P}_0/E_0$ , respectively.

The change in energy  $\Delta E_0$  can, since the momentum is given, be represented as a change  $\Delta\mu$  in rest mass of the electron. In virtue of the general relation  $E^2 = \mu^2 + P^2$ , the relation between these quantities is  $\mu\Delta\mu = E_0\Delta E_0$ . Thus we find, treating the sum of negative energies in a similar manner,

$$\begin{aligned} \Delta\mu_0 = & \frac{e^2}{2\pi^2\mu} \int d\omega d\mathbf{k} \delta(\omega^2 - k^2) \left\{ \frac{2\mu^2 - E_0 E_f + \mathbf{P}_0 \cdot \mathbf{P}_f}{E_f(E_f - E_0 + \omega)} \right. \\ & \left. + \frac{2\mu^2 + E_0 E_f + \mathbf{P}_0 \cdot \mathbf{P}_f}{E_f(E_f + E_0 + \omega)} \right\}. \end{aligned} \quad (4)$$

The integral diverges logarithmically and  $\Delta\mu_0$  defined here is meaningless. If the  $\delta(\omega^2 - k^2)$  is replaced by  $g(\omega^2 - k^2)$  defined in (1), the result is finite and invariant (i.e., does not depend on the momentum  $\mathbf{P}_0$  of the electron).

How this comes about may be seen by calculating the integral in (4) for

$$g(\omega^2 - k^2) = \delta(\omega^2 - k^2) - \delta(\omega^2 - k^2 - \lambda^2)$$

and reserving an integration on  $\lambda$  until later. The integral (4) will converge with this  $g(\omega^2 - k^2)$ , but it is convenient to divide it for purposes of calculation into the difference of two diverging ones.

This is legitimate providing the divergent integrals are first both computed over the same finite region of  $\mathbf{k}$  space, the difference taken, and then the region allowed to pass to infinity. Therefore, we shall define  $\Delta\mu_0$  by (4), in which we choose the region arbitrarily to be first over all (positive)  $\omega$  and then over a sphere in  $\mathbf{k}$  space of very large radius  $K$ . Likewise  $\Delta\mu_\lambda$  is defined as expression (4) with  $\delta(\omega^2 - k^2 - \lambda^2)$  replacing  $\delta(\omega^2 - k^2)$ , and the integration taken over the same region.

The true self-mass is therefore

$$\Delta\mu = \int_0^\infty [\text{Lim}_{K \rightarrow \infty} (\Delta\mu_0 - \Delta\mu_\lambda)] G(\lambda) d\lambda. \quad (5)$$

We may now calculate these integrals, starting with  $\Delta\mu_\lambda$ . Since  $\mathbf{P}_0 \cdot \mathbf{P}_f = \mathbf{P}_0 \cdot (\mathbf{P}_0 - \mathbf{k}) = E_0^2 - \mu^2 - \mathbf{P}_0 \cdot \mathbf{k}$  and  $E_f^2 = E_0^2 + k^2 - 2\mathbf{P}_0 \cdot \mathbf{P}_f$ , the  $\mathbf{P}_0 \cdot \mathbf{P}_f$  term in the numerator of the first term may be eliminated, the numerator becoming

$$\begin{aligned} \frac{1}{2}(E_f^2 + E_0^2 - k^2) - E_0 E_f + \mu^2 &= \mu^2 + \frac{1}{2}(\omega^2 - k^2) \\ &+ \frac{1}{2}(E_f - E_0 - \omega)(E_f - E_0 + \omega). \end{aligned}$$

Thus the first term in  $\Delta\mu_\lambda$  becomes

$$\begin{aligned} & \int \frac{\mu^2 + \frac{1}{2}(\omega^2 - k^2)}{E_f(E_f - E_0 + \omega)} \delta(\omega^2 - k^2 - \lambda^2) d\omega d\mathbf{k} \\ & + \frac{1}{2} \int \delta(\omega^2 - k^2 - \lambda^2) d\omega d\mathbf{k} (E_f - E_0 - \omega)/E_f. \end{aligned} \quad (6)$$

Adding the corresponding second term which differs from the first only in the sign of  $E_0$ , and performing the integral on  $\omega$  (which requires simply division by  $2\omega$ ), we find

$$\begin{aligned} (2\pi^2/e^2)\mu\Delta\mu_\lambda &= (\mu^2 + \frac{1}{2}\lambda^2) \int \frac{1}{(E_f + \omega)^2 - E_0^2} \\ & \cdot \frac{E_f + \omega}{E_f} \frac{d\mathbf{k}}{\omega} + \frac{1}{2} \int \frac{d\mathbf{k}}{\omega} - \frac{1}{2} \int \frac{d\mathbf{k}}{E_f}, \end{aligned} \quad (7)$$

where  $\omega = (k^2 + \lambda^2)^{\frac{1}{2}}$  and the integration is to be taken over a sphere of radius  $K$  in  $\mathbf{k}$  space. The first and, obviously, the second integrals turn out to be invariant; the third is not, but its contribution will cancel out on taking  $\Delta\mu_0 - \Delta\mu_\lambda$  as it does not depend on  $\lambda$ .<sup>7</sup> The result of the integrations<sup>8</sup> is, dropping terms of order  $1/K$  and

<sup>7</sup> Pais has suggested that one subtract from  $\Delta\mu_0$  the  $-\Delta\mu_\lambda$  that one gets not from electrodynamics but from the scalar  $f$  field (for which  $\beta \cdots \beta$  replaces  $\sum_i \alpha_i \cdots \alpha_i$ ). Proceeding in this way the integrals  $\int d\mathbf{k}/E_f$  do not appear with the same coefficient. Therefore, although this procedure leads to a finite rest mass it is not invariant in the sense here, that the limits of  $\mathbf{k}$  space integration can be taken to be independent of the momentum of the electron. A. Pais, Kon. Ned. Akad. v. Wet. Verh. D1, 19, 1 (1947).

<sup>8</sup> The first integral may be performed in the following manner: First integrate over the directions in  $\mathbf{k}$  space at constant magnitude  $k$ . Only  $E_f$  depends on the direction of  $\mathbf{k}$  and one may therefore replace the solid angle integral by one on  $E_f$ . The limits of  $E_f$  are  $E_+ = (\mu^2 + (P_0 + k)^2)^{\frac{1}{2}}$  and  $E_- = (\mu^2 + (P_0 - k)^2)^{\frac{1}{2}}$  but both terms may be considered together as one if the integral on  $k$  be extended from  $-K$  to  $K$  instead of 0 to  $K$ . To integrate this on  $k$ , substitute the variable  $x = E_+ + \omega - E_0$  and (the algebra is long) integrate by parts to reduce it to elementary integrals.

smaller:

$$\begin{aligned} (\pi/e^2)\mu\Delta\mu_\lambda &= (\mu^2 + \frac{1}{2}\lambda^2)[N_\lambda + \mu^2 X_\lambda(\mu, \mu)] \\ &\quad + \frac{1}{2}[K^2 - \lambda^2(\ln(2K/\lambda) - \frac{1}{2})] \\ &\quad - \frac{1}{2}[K^2 - \frac{1}{3}P_0^2 - \mu^2(\ln(2K/\mu) - \frac{1}{2})], \end{aligned}$$

where

$$N_\lambda = N_0 - [\lambda^2/(\lambda^2 - \mu^2)] \ln(\lambda/\mu), \quad (7a)$$

with  $N_0 = \ln(2K/\mu) - \frac{1}{2}$ , and the quantity  $X_\lambda(\mu, \mu)$  is finite as  $K \rightarrow \infty$ . It is given by setting  $\mu_0 = \mu$  in the complicated expression

$$\begin{aligned} 2\mu_0^4 X_\lambda(\mu, \mu_0) &= ((\lambda^2 - \mu^2 - \mu_0^2)^2 - 4\mu^2\mu_0^2)^{\frac{1}{2}} \\ &\quad \times \ln \frac{\lambda^2 + \mu^2 - \mu_0^2 + ((\lambda^2 - \mu^2 - \mu_0^2)^2 - 4\mu^2\mu_0^2)^{\frac{1}{2}}}{2\lambda\mu} \\ &\quad + \left( \lambda^2 - \mu^2 + \mu_0^2 - \frac{2\lambda^2\mu_0^2}{\lambda^2 - \mu^2} \right) \ln \frac{\mu}{\lambda} + \mu_0^2. \quad (7b) \end{aligned}$$

Thus  $X_0(\mu, \mu) = 1/2\mu^2$  and for  $\lambda$  large compared to  $\mu$ ,  $X_\lambda(\mu, \mu) = 1/4\lambda^2$ . Hence

$$\begin{aligned} (\pi/e^2)(\Delta\mu_0 - \Delta\mu_\lambda) &= \frac{3\mu}{2} \cdot \frac{\lambda^2}{\lambda^2 - \mu^2} \cdot \ln \frac{\lambda}{\mu} + \frac{\mu}{2} \\ &\quad - (\mu^2 + \frac{1}{2}\lambda^2)\mu X_\lambda(\mu, \mu), \quad (8) \end{aligned}$$

which is independent of  $K$  (in the limit  $K \rightarrow \infty$ ). If the important values of  $\lambda$  are much greater than  $\mu$ , we find approximately (to terms of order  $(\mu/\lambda)^2$ )

$$\Delta\mu = \mu(e^2/\pi)[\frac{3}{2} \ln(\lambda_0/\mu) + \frac{3}{8}], \quad (9)$$

where

$$\ln\lambda_0 = \int_0^\infty \ln\lambda G(\lambda) d\lambda.$$

Judging from the classical case we would have expected to take  $\lambda_0$  of order  $137\mu$ , for then all mass would be electromagnetic. But  $\Delta\mu$  here is too small for this to represent a real possibility. The experimental electron mass  $m$  is of course  $\mu + \Delta\mu$ .

The value of  $\lambda$  would have to be of phenomenal size ( $\sim e^{137}\mu$ ) before  $\Delta\mu$  can represent a sizeable fraction of the experimental mass. However, to go to the limit of the conventional electrodynamics,  $\lambda_0$  should be taken as infinite. Then the self-energy diverges logarithmically in the manner found by Weisskopf.<sup>9</sup>

The emission and subsequent absorption of a quantum acts similarly to the effect of a change in mass not only on the diagonal matrix element which we have just calculated, but on non-diagonal elements as well. Consider that the state appearing on the left of all the matrices in (3) were arbitrary, say  $x$ . Then the numerator of the first term can be expressed, as we have seen, by  $(-1/E_f)(x| -E_f + 2\beta\mu + \alpha \cdot \mathbf{P}_f | 0)$ . The second term can be expressed similarly. The two terms can be combined so that the whole expression in brackets in (3) can be written

$$2 \left\{ \frac{(x| -E_f E_0 + (E_f + \omega)(2\beta\mu + \alpha \cdot \mathbf{P}_0 - \alpha \cdot \mathbf{k}) | 0)}{E_f((E_f + \omega)^2 - E_0^2)} \right\}. \quad (10)$$

This expression may be multiplied by

$$\delta(\omega^2 - k^2 - \lambda^2)$$

and integrated with respect to  $\omega$  and over a sphere of radius  $K$  in  $\mathbf{k}$  space. We make use of the following integrals which can be directly verified:

$$\begin{aligned} \int \frac{1}{(E_f + \omega)^2 - E_0^2} \cdot \frac{E_f + \omega}{E_f} \delta(\omega^2 - k^2 - \lambda^2) d\omega d\mathbf{k} / \pi \\ = N_\lambda + \mu_0^2 X_\lambda(\mu, \mu_0), \\ \int \frac{1}{(E_f + \omega)^2 - E_0^2} \delta(\omega^2 - k^2 - \lambda^2) d\omega d\mathbf{k} / \pi \\ = \frac{1}{2} N_\lambda + \frac{1}{2} (\mu^2 + \mu_0^2 - \lambda^2) X_\lambda(\mu, \mu_0), \quad (11) \\ \int \frac{\mathbf{k}}{(E_f + \omega)^2 - E_0^2} \cdot \frac{E_f + \omega}{E_f} \delta(\omega^2 - k^2 - \lambda^2) d\omega d\mathbf{k} / \pi \\ = \frac{1}{2} \mathbf{P}_0 [\frac{1}{3} + N_\lambda + (\lambda^2 + \mu_0^2 - \mu^2) X_\lambda(\mu, \mu_0)]. \end{aligned}$$

The integrals have been calculated under the assumption that  $E_0^2 = \mu_0^2 + P_0^2$ . In our application we should take  $\mu_0 = \mu$ . The quantities  $N_\lambda$  and  $X_\lambda(\mu, \mu_0)$  are given by (7a), (7b). (The extra parameter  $\mu_0$  is helpful in obtaining other integrals, useful in the radiationless scattering problem, by differentiations with respect to the various parameters under the integral sign.)

The result of integration (10) with the density  $\delta(\omega^2 - k^2) - \delta(\omega^2 - k^2 - \lambda^2)$  is therefore

$$\begin{aligned} (\epsilon^2/\pi)(x| -E_0(\frac{1}{2}(N_0 - N_\lambda) + \frac{1}{2} - (\mu^2 - \frac{1}{2}\lambda^2)X_\lambda) \\ + (2\beta\mu + \alpha \cdot \mathbf{P}_0)(N_0 - N_\lambda + \frac{1}{2} - \mu^2 X_\lambda) \\ - \frac{1}{2}\alpha \cdot \mathbf{P}_0(N_0 - N_\lambda - \lambda^2 X_\lambda) | 0). \end{aligned}$$

<sup>9</sup> V. Weisskopf, Phys. Rev. 56, 72 (1939).

Now the energy of state 0 is  $E_0$  so that  $\alpha \cdot \mathbf{P}_0 = H_0 - \beta\mu$  is equivalent to  $E_0 - \beta\mu$ , since it operates on state 0 (no implication about state  $x$  is involved). Making this replacement, all the terms in  $E_0$  are seen to cancel and the result is simply

$$(x|\beta|0) \cdot (\Delta\mu_0 - \Delta\mu_\lambda), \quad (12)$$

where  $\Delta\mu_0 - \Delta\mu_\lambda$  is given in (8). On integrating over  $G(\lambda)d\lambda$  then we find  $(x|\beta\Delta\mu|0)$ . But this is just the perturbation element which would result from a change of mass by  $\Delta\mu$  in the Dirac equation.

We may use this result to show that the level shift for an electron in a bound state given in the present theory will be essentially that given by Weisskopf and Bethe according to their so-called wave-packet method. The change in energy of our electron in a bound state may be calculated in a straightforward manner according to the present formulation. One would simply start with Eq. (2) but with the wave functions and energies for states 0 and  $f$  being appropriate for the potential by which the electron is bound. Then one would integrate over  $g(\omega^2 - k^2)$  rather than  $\delta(\omega^2 - k^2)$  and obtain a definite finite result. The result would show a fairly large change in  $E_0$  depending logarithmically on  $\lambda$ .

A good part of this change could be accounted for as simply due to the change in  $E_0$  that would occur if the mass of the electron were altered from  $\mu$  to  $m = \mu + \Delta\mu$ . We can define the true term shift, then, as the complete change in  $E_0$ , less  $\Delta\mu(\partial E_0 / \partial \mu)$ , the change due to using  $\mu$  instead of  $m$  in computing the energy with radiation absent. But  $\partial E_0 / \partial \mu$  is by perturbation theory the expected value  $(\psi_0^* | \beta | \psi_0)$  of  $\beta$  for the state  $\psi_0$  in question. From the result (12), however, this is also equivalent to computing the self-energy of a wave packet  $\psi_0$ , assuming the electron as free. But Bethe<sup>1</sup> and Weisskopf<sup>2</sup> compute their term shift by just this prescription: the total effect less the self-energy of the free packet. The only difference here is that we would compute the term shift integral on  $g(\omega^2 - k^2)$  rather than  $\delta(\omega^2 - k^2)$ . But since the integral converges either way, the difference between the two results is very small, being of order of  $(\mu^2 / \lambda_0^2)$  times smaller than the result.

### RADIATIONLESS SCATTERING

We can study the radiationless scattering problem in a similar manner. This problem is the correction to the scattering by a first-order potential due to the possibility of emission and absorption of a virtual quantum. For example, this emission and absorption can occur at any time previous to the scattering. (It would, in this case, be nearly equivalent to a change in mass in the wave function of the electron arriving at the scatterer.) There will be a large change in cross section, which would be expected as the result of a change in mass of the electron plus a smaller change caused essentially by emissions previous to and absorptions subsequent to the scattering. As in the case of the self-energy in a field and, in fact, in all such problems, we will really be interested in those effects of radiation over and above that resulting from the change in mass. It is, therefore, simpler to compute the difference between the desired quantity calculated with no radiation and electrons of mass  $m$ , and the same quantity computed with the possibility of a virtual quantum emission and absorption with an electron of mass  $\mu$ . This difference, which we shall call the radiative correction, can be looked upon as the result of perturbation due to the addition to the Hamiltonian of both the radiative interaction terms and a term  $-\beta\Delta\mu$ . The latter term can, as we have shown, be represented by the integral over oscillators of

$$-\sum_i \left( \frac{\alpha_i \Lambda_f^+ \alpha_i}{E_f - E_0 + \omega} - \frac{\alpha_i \Lambda_f^- \alpha_i}{E_f + E_0 + \omega} \right) \quad (13)$$

when acting on a free electron state of positive energy  $E_0$  and momentum  $\mathbf{P}_0$ . When acting on a state of negative energy  $-E_0$ , the term can be shown in a similar manner to be the expression (13) with the sign of  $E_0$  changed in the denominator.

Terms like these are just the ones that Schwinger<sup>4</sup> thought should be omitted from the Hamiltonian if one wishes to get meaningful results, so that the present model agrees with Schwinger's prescription.

When this process is applied to the scattering problem to obtain the radiative correction to the matrix elements, we are left with several

residual terms. First, the emissions and absorptions previous to scattering are not exactly equivalent to a change in mass. If the emission occurs too close (in time) to the scattering, the absorption must occur in a restricted time, rather than at leisure as for a free electron forming  $\beta\Delta\mu$ . The correction to the matrix element (in the theory of holes) for this is proportional to

$$\begin{aligned} -\frac{1}{2} \sum_i \frac{(2|V\Lambda_1^+ + \alpha_i\Lambda_f^+ + \alpha_i|1)}{(E_f + \omega - E_1)^2} \\ -\frac{1}{2} \sum_i \frac{(2|V\Lambda_1^+ + \alpha_i\Lambda_f^- - \alpha_i|1)}{(E_f + \omega + E_1)^2}. \end{aligned} \quad (14)$$

We assume the potential  $V$  (vector or scalar) depending on position like  $e^{i\mathbf{q}\cdot\mathbf{R}}$  and time like  $e^{-iQt}$  induces transitions from a state 1 of momentum  $\mathbf{P}_1$ , energy  $E_1$ , to the state 2 of momentum  $\mathbf{P}_2 = \mathbf{P}_1 + \mathbf{q}$ , energy  $E_2 = E_1 + Q = (\mu^2 + P_2^2)^{\frac{1}{2}}$ . The operator  $V$  is just 1 for scalar potential,  $\alpha_x$  for vector potential in  $x$  direction, etc. The term (14) represents only that contribution due to a quantum of momentum  $\mathbf{k}$ , frequency  $\omega$ . We expect later to integrate over  $\omega$  and  $\mathbf{k}$ , times  $g(\omega^2 - k^2)$ . We put  $\mathbf{P}_f = \mathbf{P}_1 - \mathbf{k}$ ,  $E_f = (\mu^2 + P_f^2)^{\frac{1}{2}}$ . This term can also be regarded as due to the second-order normalization correction in the ordinary perturbation theory on the incoming wave function. There is a corresponding correction for the final wave function resulting from virtual quanta emitted and absorbed after the scattering: ( $\mathbf{P}_g = \mathbf{P}_2 - \mathbf{k}$ ,  $E_g = (\mu^2 + P_g^2)^{\frac{1}{2}}$ ).

$$\begin{aligned} -\frac{1}{2} \sum_i \frac{(2|\alpha_i\Lambda_g^+ + \alpha_i\Lambda_2^+ + V|1)}{(E_g + \omega - E_2)^2} \\ -\frac{1}{2} \sum_i \frac{(2|\alpha_i\Lambda_g^- - \alpha_i\Lambda_2^+ + V|1)}{(E_g + \omega + E_2)^2}. \end{aligned} \quad (15)$$

All the effects of  $\beta\Delta\mu$  are now included. The remaining terms are those for which the potential scattering occurs between the emission and absorption. They may be worked out as by Dancoff<sup>10</sup> (except that we include the longitudinal

waves by summing  $i$  from 1 to 4). They are

$$\begin{aligned} + \sum_i \frac{(2|\alpha_i\Lambda_g^+ + V\Lambda_f^+ + \alpha_i|1)}{(E_f + \omega - E_1)(E_g + \omega - E_2)} \\ + \sum_i \frac{(2|\alpha_i\Lambda_g^- - V\Lambda_f^- - \alpha_i|1)}{(E_f + \omega + E_1)(E_g + \omega + E_2)} \end{aligned} \quad (16)$$

and

$$\begin{aligned} - \sum_i \frac{(2|\alpha_i\Lambda_g^+ + V\Lambda_f^- - \alpha_i|1)}{(E_g + \omega - E_2)(E_f + \omega + E_1)} \\ \times \left[ 1 + \frac{2\omega}{E_f + E_g - E_2 + E_1} \right] \\ - \sum_i \frac{(2|\alpha_i\Lambda_g^- - V\Lambda_f^+ + \alpha_i|1)}{(E_f + \omega - E_1)(E_g + \omega + E_2)} \\ \times \left[ 1 + \frac{2\omega}{E_f + E_g + E_2 - E_1} \right]. \end{aligned} \quad (17)$$

Although each separate term diverges, the sum of (14), (15), (16), (17) will lead to an integral convergent for large  $\mathbf{k}$  even if integrated in the conventional manner on  $\delta(\omega^2 - k^2)$ . This is the result of Lewis. Integration on  $g(\omega^2 - k^2)$  will make each term converge for large  $\mathbf{k}$ , but will then only make correction to the sum of order  $(\mu/\lambda)^2$  smaller. These we shall neglect.

The integrals do, however, diverge logarithmically at the lower limit of small momentum transfer. This infra-red catastrophe has been completely cleared up by Bloch and Nordsieck.<sup>11</sup> They show that for very long wave-length quanta the amplitude for emission and reabsorption of more than one quantum is not negligible. Inclusion of these higher order terms, which is necessary only in the non-relativistic region, solves the problem. To keep the results given here in a simple form, we can imagine the integrals to be performed down to some minimum momentum  $k_{\min}$ , small compared to  $\mu$ . What is effectively the same thing but which is easier (because relativistic invariance is maintained) for practical purposes, is to imagine that the quanta have a very small rest mass  $\lambda_{\min}$ . Thus we integrate the density

$$\delta(\omega^2 - k^2 - \lambda_{\min}^2) d\omega dk$$

<sup>10</sup> S. M. Dancoff, Phys. Rev. 55, 959 (1939).

<sup>11</sup> F. Bloch and A. Nordsieck, Phys. Rev. 52, 54 (1937).

and assume  $\lambda_{\min} \ll \mu$ . The two methods are equivalent if one replaces  $\ln \lambda_{\min}$  by  $\ln(2k_{\min}) - 1$ .

The integrals may be expanded in powers of  $q$  and  $Q$ , say up to the second.<sup>12</sup> The constant term vanishes on integration. The integrals appearing may all be expressed in terms of various parametric derivatives of the integrals already given in (11). The result may be expressed in terms of a general potential in a very simple way. A term linear in  $q$ , such as proportional to  $q_x$  say, is equivalent to taking the matrix element  $(2|q_x \exp(iq \cdot R)|1)$  directly between the two states 2, 1. But this is also equivalent to the matrix element of  $-i(\partial/\partial x) \exp(iq \cdot R)$ . Thus if the potential varied in any other manner in space, one sees by superposition that the matrix element is the same as that of  $-i\partial V/\partial x$ . Thus the terms up to second order can be represented by matrix elements of first and second space and time derivatives of the potential. That is, the radiative correction to the scattering in any potential is equivalent to the first order in  $e^2$  and in the potential, to the scattering produced by a perturbation  $\Delta H$  to the Dirac Hamiltonian. The perturbation up to terms of first and second derivatives of the vector potential  $A$  and the scalar potential  $\varphi$  is calculated in this manner to be

$$\begin{aligned} \Delta H = & \frac{e^2}{2\pi\hbar c} \left\{ -\frac{\hbar e}{2\mu c} (\beta(\sigma \cdot B) - i\beta\alpha \cdot E) \right. \\ & \left. + \frac{2\hbar^2 e}{3\mu^2 c^2} (\square^2 \varphi - \alpha \cdot \square^2 A) \left( \ln \frac{\mu}{\lambda_{\min}} - \frac{3}{8} \right) \right\}. \quad (18) \end{aligned}$$

The first term, where  $B = \nabla \times A$  and  $E = -\nabla \varphi - (1/c)\partial A / \partial t$ , has the same effect as an alteration in the electron magnetic moment<sup>13</sup> by a fraction  $e^2/2\pi\hbar c$ . This effect was first discovered by Schwinger.<sup>4</sup>

#### LINE SHIFT

The perturbation to  $H$  given here is useful not only for scattering problems but also for the line-shift problem. The actual motion of an electron in a binding potential can be visualized

<sup>12</sup> The integrals have also been worked out, by other methods, for arbitrarily large  $q$  and  $Q$ . These will appear in a future publication.

<sup>13</sup> W. Pauli, *Handbuch der Physik* (1933), Vol. 24/1, p. 233.

as simply a continued sequence of scatterings in this potential. For each scattering we can calculate the effect of virtual quanta in the way outlined above. However, it is possible, if the potential is strong, that *two* scatterings occur between the emission and reabsorption of the quantum, in which case the above formula for  $\Delta H$  is incorrect. In hydrogen the potential over most of the atom is sufficiently weak that this does not occur with effective probability. The very long wave-length quanta do have a tendency to exist in the virtual state for long periods, but they have been eliminated by the cut-off  $\lambda_{\min}$  at low frequencies.

In hydrogen, then, the line shift due to quanta above minimum wave number  $k_{\min}$  is the expected value, for the state in question, of

$$\begin{aligned} \Delta H = & \frac{e^2}{2\pi\hbar c} \left\{ -\frac{\hbar e i}{2\mu c} \beta \alpha \cdot \nabla \varphi + \frac{2\hbar^2 e}{3\mu^2 c^2} (\nabla^2 \varphi) \right. \\ & \left. \times \left( \ln \frac{\mu c}{2\hbar k_{\min}} + \frac{5}{8} \right) \right\}, \quad (19) \end{aligned}$$

where  $\varphi = e/r$ ,  $r$  being the distance to the proton, and we have used the relation

$$\ln \lambda_{\min} = \ln(2k_{\min}) - 1.$$

The first term insures that the fine structure separation correction will be that expected from the change in the electron's magnetic moment. The second may be combined with Bethe's non-relativistic calculation for quanta below  $k_{\min}$ .<sup>14</sup>

#### APPLICATION TO OTHER PROCESSES

The important problem of verifying that the self-energy will not diverge in higher-order approximations has not been carried to completion. It appears unlikely that trouble will arise here. If that is true the model probably gives sensible answers to all problems of quantum electrodynamics other than those involving Uehling polarization effects, discussed below. It has been found to give finite self-mass if we have, instead of a vector field, a scalar field or a pseudoscalar field, coupled to the electron in the simplest way possible without gradient operators. If the field

<sup>14</sup> Using Eq. (18), Professor Bethe finds 1050 megacycles for the separation between  $2p_{3/2}$  and  $2s_{1/2}$  in hydrogen. (Solvay Report.)

quanta have mass  $M$ ,  $g(\omega^2 - k^2)$  is replaced by  $g(\omega^2 - k^2 - M^2)$ , and the values of  $\lambda$  of importance are chosen to be large compared to  $M$ .

The results for electrodynamics, then, after mass renormalization, depend only slightly on the form of  $G(\lambda)$  and the size of  $\lambda_0$ . Since  $\lambda_0$  may be taken to be extremely large without spoiling the smallness of  $\Delta\mu$ , there would appear to be good reason to drop the dependence on  $\lambda$  altogether. Thus the  $G(\lambda)$  appears only as a complicated scaffold which is removed after the calculation is done.

On the other hand, electrodynamics probably does break down somewhere and it is interesting to keep the terms in  $\lambda$  for various phenomena to see if one might be selected which is particularly sensitive to  $\lambda$ . This phenomena would then be a promising one to study experimentally. The Møller interaction between two electrons is modified by the present theory. There is, of course, the radiative correction, but in addition to that there is simply a change due to the change in the density function for the quanta which can be exchanged. The Møller interaction ordinarily is proportional to  $1/q^2$ , where  $q$  is the magnitude of the momentum transferred from one electron to the other in the center of gravity system. The modification is only that this factor is changed to  $f_0^2(1/q^2 - 1/(q^2 + \lambda^2))G(\lambda)d\lambda$ . This represents a decrease in cross section for hard collisions. If  $\lambda$  is of order  $137 \mu c^2$ , we would need electrons in the center of gravity system of roughly 30 Mev to find a strong effect. This corresponds, however, to bombardment of stationary electrons by electrons of  $3\frac{1}{2}$  Bev.<sup>16</sup>

It is interesting to note that the Møller interaction can be viewed as simply a correction to self-energy due to the exclusion principle. The self-energy of two electrons, 1 and 2, is not the sum of the self-energy of each, for one of the virtual states that 2 could ordinarily enter by emission of a quantum is now occupied by 1. The difference between the self-energy of two electrons and the sum of the self-energy of each

separately comes out to be just their interaction energy.

#### VACUUM POLARIZATION. ALTERNATIVE CUT-OFF PROCEDURES

In the above calculation, terms of the type discussed by Uehling<sup>16</sup> have been omitted. These terms represent processes involving a pair production followed by annihilation of the same pair. For example, a pair produced by the potential may annihilate again emitting a quantum. This quantum is then absorbed by the electron in state 1 transferring it to state 2. These terms are infinite and are not made convergent by the present scheme. There is some point, nevertheless, to solving problems at first without taking them into account. This is because their net effect is only to alter the effective potential in which the electron finds itself, for it may be scattered either directly or by the quantum produced by the Uehling terms. That is, if this problem of polarization of the vacuum is solved it will mean, if there is any effect, simply that the potential  $A$ ,  $\varphi$  appearing in the Dirac equation and (to high order) in such terms as (18) should be replaced by new "polarized" potentials  $A'$ ,  $\varphi'$ .

These polarization terms can be characterized in a relativistically invariant manner. All the terms which have been calculated above contain matrix elements of operators between states in a sequence such as 1 to  $f$ ,  $f$  to  $g$ ,  $g$  to 2. The omitted polarization terms contain transitions like  $f$  to  $g$ ,  $g$  to  $f$ , 1 to 2. For higher order processes the polarization terms are those which do not contain a continued sequence of transitions from the initial to the final state.

The polarization terms are not affected in any helpful way by the changes in the density of quanta. It is likely that this problem will have its answer in a changed physical viewpoint. However, there is a simple alternative procedure to produce finite self-energies which also makes convergent the integrals appearing in Serber's<sup>17</sup> treatment of the polarization problem. (Since, however, this treatment of Serber already presupposes a partial subtraction procedure of Heisenberg and Dirac, the situation is not so clear here as in the self-energy problem.)

<sup>16</sup> A more promising way to obtain processes with high momentum transfer would be wide-angle scattering of electrons from nuclei. But here deviations from expectations might be associated with uncertainties in the nuclear charge distributions rather than electrodynamics. Very wide angle pair production is a phenomena which does occur for high energy incident  $\gamma$ -rays with large momentum transfer in a region not too close to the nucleus.

<sup>17</sup> E. A. Uehling, Phys. Rev. **48**, 55 (1935).

<sup>17</sup> R. Serber, Phys. Rev. **48**, 49 (1935).

From the point of view of coordinate space, the reason that the electronic self-energy diverges appears to be this. A virtual light quantum emitted at one point spreads out as  $\delta(t^2 - r^2)$  from the origin. The wave packet of the electron spreading out after the emission of the quantum has, as a consequence of Dirac's equation, a similar discontinuous value along the light cone. It is the continued coincidence of these singularities which makes the matrix element for the subsequent absorption of the quantum infinite. The method outlined above of changing  $\delta(\omega^2 - k^2)$  to  $g(\omega^2 - k^2)$  has the effect of changing  $\delta(t^2 - r^2)$  to  $f(t^2 - r^2)$  where  $f(s^2)$  is everywhere finite and goes to zero rapidly for  $|s^2| > 1/\lambda_0^2$ . The quanta have been moved away from the electrons so that overlap on the light cone is reduced.

An obvious alternative procedure is to move the electron wave function away from the quanta. This is easily done in a very similar manner. We assume the density of electron states of energy  $E$ , momentum  $\mathbf{P}$  to be  $g(E^2 - P^2 - \mu^2)$  rather than  $\delta(E^2 - P^2 - \mu^2)$ .<sup>18</sup> The quanta are conventional,  $\omega = k$ , density  $d\mathbf{k}/k$ . The self-energy integrals (2) can, of course, be expressed as an integral over the intermediate state momentum  $\mathbf{P}_f$  rather than  $\mathbf{k}$ . Replacing  $d\mathbf{P}_f/E_f$  by  $g(E_f^2 - P_f^2 - \mu^2)dE_f d\mathbf{P}_f$ , we find

$$\Delta E'_0 = -\frac{e^2}{2\pi^2} \int g(E_f^2 - P_f^2 - \mu^2) dE_f d\mathbf{P}_f \cdot \frac{E_f}{k} \sum_i \left\{ \frac{(0|\alpha_i \Lambda_f^+ \alpha_i|0)}{E_f + k - E_0} - \frac{(0|\alpha_i \Lambda_f^- \alpha_i|0)}{E_f + k + E_0} \right\},$$

<sup>18</sup> This is seen to be essentially the method proposed by Wataghin. G. Wataghin, Zeits. f. Physik **88**, 92 (1934).

where  $k = |\mathbf{P}_f - \mathbf{P}_0|$ ,  $E_0 = (\mu^2 + P_0^2)^{1/2}$ . The projection operators are unchanged since it is only the density of states which we wish to alter. They are still  $\Lambda_f^{2\pm} = (E_f \pm \alpha \cdot \mathbf{P}_f \pm \beta \mu)/2E_f$ . The result of this calculation is to verify that  $\Delta E'_0$  is finite, (depending logarithmically on  $\lambda_0$ ). The other problems can be analyzed in the same way.

In the problem of polarization of the vacuum, the wave functions of both electron and positron ordinarily spread with a singularity on the light cone. The matrix element for their subsequent annihilation is therefore infinite. With the modification here described these wave functions are made less singular and their overlap integral is finite. The polarization integrals in Serber's article<sup>17</sup> may now be integrated to yield finite results.

Other than terms which might be removed by a small renormalization of charge (depending logarithmically on  $\lambda_0$ ), the net effect in (17) would be to change the  $-(\frac{3}{8})$  in the last term of (17) to  $-(\frac{3}{8}) - (\frac{1}{5})$ . However, the real existence of such polarization corrections is, in the author's view, uncertain. These matters will be discussed in much more detail in future publications. Also reserved for future publication is a more complete physical theory from which the results reported here may be directly deduced. It yields much more powerful techniques for setting up problems and performing the required integrations.

The author would like to express his gratitude to Mr. P. V. C. Hough for assistance in the calculations and to Professor H. A. Bethe and Dr. F. Dyson and many others for useful discussions.



## Space-Time Approach to Quantum Electrodynamics

R.P. Feynman

Department of Physics, Cornell University,

Ithaca, New York

(Received May 9, 1949)



Reprinted in "Quantum Electrodynamics", edited by Julian Schwinger



### Abstract

In this paper two things are done. (1) It is shown that a considerable simplification can be attained in writing down matrix elements for complex processes in electrodynamics. Further, a physical point of view is available which permits them to be written down directly for any specific problem. Being simply a restatement of conventional electrodynamics, however, the matrix elements diverge for complex processes. (2) Electrodynamics is modified by altering the interaction of electrons at short distances. All matrix elements are now finite, with the exception of those relating to problems of vacuum polarization. The latter are evaluated in a manner suggested by Pauli and Bethe, which gives finite results for these matrices also. The only effects sensitive to the modification are changes in mass and charge of the electrons. Such changes could not be directly observed. Phenomena directly observable, are insensitive to the details of the modification used (except at extreme energies). For such phenomena, a limit can be taken as the range of the modification goes to zero. The results then agree with those of Schwinger. A complete, unambiguous, and presumably consistent, method is therefore available for the calculation of all processes involving electrons and photons.

The simplification in writing the expressions results from an emphasis on the over-all space-time view resulting from a study of the solution of the equations of electrodynamics. The relation of this to the more conventional Hamiltonian point of view is discussed. It would be very difficult to make the modification which is proposed if one insisted on having the equations in Hamiltonian form.

The methods apply as well to charges obeying the Klein-Gordon equation, and to the various meson theories of nuclear forces. Illustrative examples are given. Although a modification like that used in electrodynamics can make all matrices finite for all of the meson theories, for some of the theories it is no longer true that all directly observable phenomena are insensitive to the details of the modification used.

The actual evaluation of integrals appearing in the matrix elements may be facilitated, in the simpler cases, by methods described in the appendix.

This paper should be considered as a direct continuation of a preceding one<sup>1</sup> (I) in which the motion of electrons, neglecting interaction, was analyzed, by dealing directly with the *solution* of the Hamiltonian differential equations. Here the same technique is applied to include interactions and in that way to express in simple terms the solution of problems in quantum electrodynamics.

For most practical calculations in quantum electrodynamics the solution is ordinarily expressed in terms of a matrix element. The matrix is worked out as an expansion in powers of  $e^2/\hbar c$ , the successive terms corresponding to the inclusion of an increasing number of virtual quanta. It appears that a considerable simplification can be achieved in writing down these matrix elements for complex processes. Furthermore, each term in the expansion can be written down and understood directly from a physical point of view, similar to the space-time view in I. It is the purpose of this paper to describe how this may be done. We shall also discuss methods of handling the divergent integrals which appear in these matrix elements.

The simplification in the formulae results mainly from the fact that previous methods unnecessarily separated into individual terms processes that were closely related physically. For example, in the exchange of a quantum between two electrons there were two terms depending on which electron emitted and which absorbed the quantum. Yet, in the virtual states considered, timing relations are not significant. Only the order of operators in the matrix must be maintained. We have seen (I), that in addition, processes in

---

<sup>1</sup>R. P. Feynman, Phys. Rev. **76**, 749 (1949), hereafter called I.

which virtual pairs are produced can be combined with others in which only positive energy electrons are involved. Further, the effects of longitudinal and transverse waves can be combined together. The separations previously made were on an unrelativistic basis (reflected in the circumstance that apparently momentum but not energy is conserved in intermediate states). When the terms are combined and simplified, the relativistic invariance of the result is self-evident.

We begin by discussing the solution in space and time of the Schrödinger equation for particles interacting instantaneously. The results are immediately generalizable to delayed interactions of relativistic electrons and we represent in that way the laws of quantum electrodynamics. We can then see how the matrix element for any process can be written down directly. In particular, the self-energy expression is written down.

So far, nothing has been done other than a restatement of conventional electrodynamics in other terms. Therefore, the self-energy diverges. A modification<sup>2</sup> in interaction between charges is next made, and it is shown that the self-energy is made convergent and corresponds to a correction to the electron mass. After the mass correction is made, other real processes are finite and insensitive to the “width” of the cut-off in the interaction.<sup>3</sup>

Unfortunately, the modification proposed is not completely satisfactory theoretically (it leads to some difficulties of conservation of energy). It does, however, seem consistent and satisfactory to define the matrix element for all real processes as the limit of that computed here as the cut-off width goes to zero. A similar technique suggested by Pauli and by Bethe can be applied to problems of vacuum polarization (resulting in a renormalization of charge) but again a strict physical basis for the rules of convergence is not known.

After mass and charge renormalization, the limit of zero cut-off width can be taken for all real processes. The results are then equivalent to those of Schwinger<sup>4</sup> who does not make explicit use of the convergence factors. The method of Schwinger is to identify the terms corresponding to corrections in mass and charge and, previous to their evaluation, to remove them from the expressions for real processes. This has the advantage of showing that the results can be strictly independent of particular cut-off methods.

---

<sup>2</sup>For a discussion of this modification in classical physics see R. P. Feynman, Phys. Rev. **74**, 939 (1948), hereafter referred to as A.

<sup>3</sup>A brief summary of the methods and results will be found in R. P. Feynman, Phys. Rev. **74**, 1430 (1948), hereafter referred to as B.

<sup>4</sup>J. Schwinger, Phys. Rev. **74**, 1439 (1948), Phys. Rev. **75**, 651 (1949). A proof of this equivalence is given by F. J. Dyson, Phys. Rev. **75**, 486 (1949).

On the other hand, many of the properties of the integrals are analyzed using formal properties of invariant propagation functions. But one of the properties is that the integrals are infinite and it is not clear to what extent this invalidates the demonstrations. A practical advantage of the present method is that ambiguities can be more easily resolved; simply by direct calculation of the otherwise divergent integrals. Nevertheless, it is not at all clear that the convergence factors do not upset the physical consistency of the theory. Although in the limit the two methods agree, neither method appears to be thoroughly satisfactory theoretically. Nevertheless, it does appear that we now have available a complete and definite method for the calculation of physical processes to any order in quantum electrodynamics.

Since we can write down the solution to any physical problem, we have a complete theory which could stand by itself. It will be theoretically incomplete, however, in two respects. First, although each term of increasing order in  $e^2/\hbar c$  can be written down it would be desirable to see some way of expressing things in finite form to all orders in  $e^2/\hbar c$  at once. Second, although it will be physically evident that the results obtained are equivalent to those obtained by conventional electrodynamics the mathematical proof of this is not included. Both of these limitations will be removed in a subsequent paper (see also Dyson<sup>5</sup>).

Briefly the genesis of this theory was this. The conventional electrodynamics was expressed in the Lagrangian form of quantum mechanics described in the Reviews of Modern Physics.<sup>5</sup> The motion of the field oscillators could be integrated out (as described in Section 13 of that paper), the result being an expression of the delayed interaction of the particles. Next the modification of the delta-function interaction could be made directly from the analogy to the classical case.<sup>6</sup> This was still not complete because the Lagrangian method had been worked out in detail only for particles obeying the non-relativistic Schrödinger equation. It was then modified in accordance with the requirements of the Dirac equation and the phenomenon of pair creation. This was made easier by the reinterpretation of the theory of holes (I). Finally for practical calculations the expressions were developed in a power series in  $e^2/\hbar c$ . It was apparent that each term in the series had a simple physical interpretation. Since the result was easier to understand than the derivation, it was thought best to publish the results first in this

---

<sup>5</sup>R. P. Feynman, Rev. Mod. Phys. **20**, 367 (1948). The application to electrodynamics is described in detail by H. J. Groenewold, Koninklijke Nederlandsche Akademie van Wetenschappen. Proceedings Vol. LII, **3** (226) 1949.

<sup>6</sup>For a discussion of this modification in classical physics see R. P. Feynman, Phys. Rev. **74** 939 (1948), hereafter referred to as A.

paper. Considerable time has been spent to make these first two papers as complete and as physically plausible as possible without relying on the Lagrangian method, because it is not generally familiar. It is realized that such a description cannot carry the conviction of truth which would accompany the derivation. On the other hand, in the interest of keeping simple things simple the derivation will appear in a separate paper.

The possible application of these methods to the various meson theories is discussed briefly. The formulas corresponding to a charge particle of zero spin moving in accordance with the Klein Gordon equation are also given. In an Appendix a method is given for calculating the integrals appearing in the matrix elements for the simpler processes.

The point of view which is taken here of the interaction of charges differs from the more usual point of view of field theory. Furthermore, the familiar Hamiltonian form of quantum mechanics must be compared to the overall space-time view used here. The first section is, therefore, devoted to a discussion of the relations of these viewpoints.

## 1 COMPARISON WITH THE HAMILTONIAN METHOD

Electrodynamics can be looked upon in two equivalent and complementary ways. One is as the description of the behavior of a field (Maxwell's equations). The other is as a description of a direct interaction at a distance (albeit delayed in time) between charges (the solutions of Lienard and Wiechert). From the latter point of view light is considered as an interaction of the charges in the source with those in the absorber. This is an impractical point of view because many kinds of sources produce the same kind of effects. The field point of view separates these aspects into two simpler problems, production of light, and absorption of light. On the other hand, the field point of view is less practical when dealing with close collisions of particles (or their action on themselves). For here the source and absorber are not readily distinguishable, there is an intimate exchange of quanta. The fields are so closely determined by the motions of the particles that it is just as well not to separate the question into two problems but to consider the process as a direct interaction. Roughly, the field point of view is most practical for problems involving real quanta, while the interaction view is best for the discussion of the virtual quanta involved. We shall emphasize the interaction viewpoint in this paper, first because it is less familiar and therefore requires more discussion, and second because the important aspect

in the problems with which we shall deal is the effect of virtual quanta.

The Hamiltonian method is not well adapted to represent the direct action at a distance between charges because that action is delayed. The Hamiltonian method represents the future as developing out of the present. If the values of a complete set of quantities are known now, their values can be computed at the next instant in time. If particles interact through a delayed interaction, however, one cannot predict the future by simply knowing the present motion of the particles. One would also have to know what the motions of the particles were in the past in view of the interaction this may have on the future motions. This is done in the Hamiltonian electrodynamics, of course, by requiring that one specify besides the present motion of the particles, the values of a host of new variables (the coordinates of the field oscillators) to keep track of that aspect of the past motions of the particles which determines their future behavior. The use of the Hamiltonian forces one to choose the field viewpoint rather than the interaction viewpoint.

In many problems, for example, the close collisions of particles, we are not interested in the precise temporal sequence of events. It is not of interest to be able to say how the situation would look at each instant of time during a collision and how it progresses from instant to instant. Such ideas are only useful for events taking a long time and for which we can readily obtain information during the intervening period. For collisions it is much easier to treat the process as a whole.<sup>7</sup> The Møller interaction matrix for the collision of two electrons is not essentially more complicated than the nonrelativistic Rutherford formula, yet the mathematical machinery used to obtain the former from quantum electrodynamics is vastly more complicated than Schrödinger's equation with the  $e^2/r_{12}$  interaction needed to obtain the latter. The difference is only that in the latter the action is instantaneous so that the Hamiltonian method requires no extra variables, while in the former relativistic case it is delayed and the Hamiltonian method is very cumbersome.

We shall be discussing the solutions of equations rather than the time differential equations from which they come. We shall discover that the solutions, because of the over-all space-time view that they permit, are as easy to understand when interactions are delayed as when they are instantaneous.

As a further point, relativistic invariance will be self-evident. The Hamiltonian form of the equations develops the future from the instantaneous present. But for different observers in relative motion the instantaneous present is different, and corresponds to a different 3-dimensional cut of space-

---

<sup>7</sup>This is the viewpoint of the theory of the  $S$  matrix of Heisenberg.

time. Thus the temporal analyses of different observers is different and their Hamiltonian equations are developing the process in different ways. These differences are irrelevant, however, for the solution is the same in any space time frame. By forsaking the Hamiltonian method, the wedding of relativity and quantum mechanics can be accomplished most naturally.

We illustrate these points in the next section by studying the solution of Schrödinger's equation for non-relativistic particles interacting by an instantaneous Coulomb potential (Eq. 2). When the solution is modified to include the effects of delay in the interaction and the relativistic properties of the electrons we obtain an expression of the laws of quantum electrodynamics (Eq. 4).

## 2 THE INTERACTION BETWEEN CHARGES

We study by the same methods as in I, the interaction of two particles using the same notation as I. We start by considering the non-relativistic case described by the Schrödinger equation (I, Eq. 1). The wave function at a given time is a function  $\psi(\mathbf{x}_a, \mathbf{x}_b, t)$  of the coordinates  $\mathbf{x}_a$  and  $\mathbf{x}_b$  of each particle. Thus call  $K(\mathbf{x}_a, \mathbf{x}_b, t; \mathbf{x}'_a, \mathbf{x}'_b, t')$  the amplitude that particle  $a$  at  $\mathbf{x}'_a$  at time  $t'$  will get to  $\mathbf{x}_a$  at  $t$  while particle  $b$  at  $\mathbf{x}'_b$  at  $t'$  gets to  $\mathbf{x}_b$  at  $t$ . If the particles are free and do not interact this is

$$K(\mathbf{x}_a, \mathbf{x}_b, t; \mathbf{x}'_a, \mathbf{x}'_b, t') = K_{0a}(\mathbf{x}_a, t; \mathbf{x}'_a, t')K_{0b}(\mathbf{x}_b, t; \mathbf{x}'_b, t')$$

where  $K_{0a}$  is the  $K_0$  function for particle  $a$  considered as free. In *this* case we can obviously define a quantity like  $K$ , but for which the time  $t$  need not be the same for particles  $a$  and  $b$  (likewise for  $t'$ ); e.g.,

$$K_0(3, 4; 1, 2) = K_{0a}(3, 1)K_{0b}(4, 2) \quad (1)$$

can be thought of as the amplitude that particle  $a$  goes from  $\mathbf{x}_1$  at  $t_1$  to  $\mathbf{x}_3$  at  $t_3$  and that particle  $b$  goes from  $\mathbf{x}_2$  at  $t_2$  to  $\mathbf{x}_4$  at  $t_4$ .

When the particles do interact, one can only define the quantity  $K(3, 4; 1, 2)$  precisely if the interaction vanishes between  $t_1$  and  $t_2$  and also between  $t_3$  and  $t_4$ . In a real physical system such is not the case. There is such an enormous advantage, however, to the concept that we shall continue to use it, imagining that we can neglect the effect of interactions between  $t_1$  and  $t_2$  and between  $t_3$  and  $t_4$ . For practical problems this means choosing such long time intervals  $t_3 - t_1$  and  $t_4 - t_2$  that the extra interactions near the end points have small relative effects. As an example, in a scattering

problem it may well be that the particles are so well separated initially and finally that the interaction at these times is negligible. Again energy values can be defined by the average rate of change of phase over such long time intervals that errors initially and finally can be neglected. Inasmuch as any physical problem can be defined in terms of scattering processes we do not lose much in a general theoretical sense by this approximation. If it is not

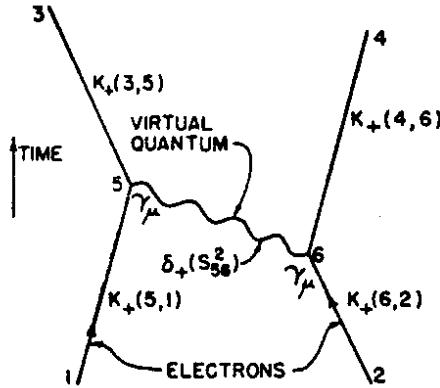


Figure 1: The fundamental interaction Eq. (4). Exchange of one quantum between two electrons.

made it is not easy to study interacting particles relativistically, for there is nothing significant in choosing  $t_1 = t_3$  if  $\mathbf{x}_1 \neq \mathbf{x}_3$  as absolute simultaneity of events at a distance cannot be defined invariantly. It is essentially to avoid this approximation that the complicated structure of the older quantum electrodynamics has been built up. We wish to describe electrodynamics as a delayed interaction between particles. If we can make the approximation of assuming a meaning to  $K(3, 4; 1, 2)$  the results of this interaction can be expressed very simply.

To see how this may be done, imagine first that the interaction is simply that given by a Coulomb potential  $e^2/r$  where  $r$  is the distance between the particles. If this be turned on only for a very short time  $\Delta t_0$  at time  $t_0$  the first order correction to  $K(3, 4; 1, 2)$  can be worked out exactly as was Eq. (9) of I by an obvious generalization to two particles:

$$K^{(1)}(3, 4; 1, 2) = -ie^2 \int \int K_{0a}(3, 5) K_{0b}(4, 6) r_{56}^{-1} \\ \times K_{0a}(5, 1) K_{0b}(6, 2) d^3 \mathbf{x}_5 d^3 \mathbf{x}_6 \Delta t_0,$$

where  $t_5 = t_6 = t_0$ . If now the potential were on at all times (so that strictly  $K$  is not defined unless  $t_4 = t_3$  and  $t_1 = t_2$ ), the first-order effect is obtained by integrating on  $t_0$ , which we can write as an integral over both  $t_5$  and  $t_6$  if we include a delta-function  $\delta(t_5 - t_6)$  to insure contribution only when  $t_5 = t_6$ . Hence, the first-order effect of interaction is (calling  $t_5 - t_6 = t_{56}$ ):

$$\begin{aligned} K^{(1)}(3, 4; 1, 2) &= -ie^2 \int \int K_{0a}(3, 5) K_{0b}(4, 6) r_{56}^{-1} \\ &\quad \times \delta(t_{56}) K_{0a}(5, 1) K_{0b}(6, 2) d\tau_5 d\tau_6, \end{aligned} \tag{2}$$

where  $d\tau = d^3 \mathbf{x} dt$ .

We know, however, in classical electrodynamics, that the Coulomb potential does not act instantaneously, but is delayed by a time  $r_{56}$ , taking the speed of light as unity. This suggests simply replacing  $r_{56}^{-1} \delta(t_{56})$  in (2) by something like  $r_{56}^{-1} \delta(t_{56} - r_{56})$  to represent the delay in the effect of  $b$  on  $a$ .

This turns out to be not quite right,<sup>8</sup> for when this interaction is represented by photons they must be of only positive energy, while the Fourier transform of  $\delta(t_{56} - r_{56})$  contains frequencies of both signs. It should instead be replaced by  $\delta_+(t_{56} - r_{56})$  where

$$\delta_+(x) = \int_0^\infty e^{-i\omega x} d\omega / \pi = \lim_{\epsilon \rightarrow 0} \frac{(\pi i)^{-1}}{x - i\epsilon} = \delta(x) + (\pi ix)^{-1}. \tag{3}$$

This is to be averaged with  $r_{56}^{-1} \delta_+(-t_{56} - r_{56})$  which arises when  $t_5 < t_6$  and corresponds to  $a$  emitting the quantum which  $b$  receives. Since

$$(2r)^{-1} (\delta_+(t - r) + \delta_+(-t - r)) = \delta_+(t^2 - r^2),$$

this means  $r_{56}^{-1} \delta(t_{56})$  is replaced by  $\delta_+(s_{56}^2)$  where  $s_{56}^2 = t_{56}^2 - r_{56}^2$  is the square of the relativistically invariant interval between points 5 and 6. Since in classical electrodynamics there is also an interaction through the vector potential, the complete interaction (see A, Eq. (I)) should be  $(1 - (\mathbf{v}_5 \cdot \mathbf{v}_6)) \delta_+(s_{56}^2)$ , or in the relativistic case,

$$(1 - \alpha_a \cdot \alpha_b) \delta_+(s_{56}^2) = \beta_a \beta_b \gamma_{a\mu} \gamma_{b\mu} \delta_+(s_{56}^2).$$

---

<sup>8</sup>It and a like term for the effect of  $a$  on  $b$ , leads to a theory which, in the classical limit, exhibits interaction through half-advanced and half-retarded potentials. Classically, this is equivalent to purely retarded effects within a closed box from which no light escapes (e.g., see A, or J. A. Wheeler and R. P. Feynman, Rev. Mod. Phys. **17**, 157 (1945)). Analogous theorems exist in quantum mechanics but it would lead us too far astray to discuss them now.

Hence we have for electrons obeying the Dirac equation,

$$K^{(1)}(3, 4; 1, 2) = ie^2 \int \int K_{+a}(3, 5) K_{+b}(4, 6) \gamma_{a\mu} \gamma_{b\mu} \\ \times \delta_+(s_{56}^2) K_{+a}(5, 1) K_{+b}(6, 2) d\tau_5 d\tau_6, \quad (4)$$

where  $\gamma_{a\mu}$  and  $\gamma_{b\mu}$ , are the Dirac matrices applying to the spinor corresponding to particles  $a$  and  $b$ , respectively (the factor  $\beta_a \beta_b$  being absorbed in the definition, I Eq. (17), of  $K_+$ ).

This is our fundamental equation for electrodynamics. It describes the effect of exchange of one quantum (therefore first order in  $e^2$ ) between two electrons. It will serve as a prototype enabling us to write down the corresponding quantities involving the exchange of two or more quanta between two electrons or the interaction of an electron with itself. It is a consequence of conventional electrodynamics. Relativistic invariance is clear. Since one sums over  $\mu$  it contains the effects of both longitudinal and transverse waves in a relativistically symmetrical way.

We shall now interpret Eq. (4) in a manner which will permit us to write down the higher order terms. It can be understood (see Fig. 1) as saying that the amplitude for “ $a$ ” to go from 1 to 3 and “ $b$ ” to go from 2 to 4 is altered to first order because they can exchange a quantum. Thus, “ $a$ ” can go to 5 (amplitude  $(K_+(5, 1))$ ) emit a quantum (longitudinal, transverse, or scalar  $\gamma_{a\mu}$ ) and then proceed to 3 ( $K_+(3, 5)$ ). Meantime “ $b$ ” goes to 6 ( $K_+(6, 2)$ ), absorbs the quantum ( $\gamma_{b\mu}$ ) and proceeds to 4 ( $K_+(4, 6)$ ). The quantum meanwhile proceeds from 5 to 6, which it does with amplitude  $\delta_+(s_{56}^2)$ . We must sum over all the possible quantum polarizations it and positions and times of emission 5, and of absorption 6. Actually if  $t_5 > t_6$  it would be better to say that “ $a$ ” absorbs and “ $b$ ” emits but no attention need be paid to these matters, as all such alternatives are automatically contained in (4).

The correct terms of higher order in  $e^2$  or involving larger numbers of electrons (interacting with themselves or in pairs) can be written down by the same kind of reasoning. They will be illustrated by examples as we proceed. In a succeeding paper they will all be deduced from conventional quantum electrodynamics.

Calculation, from (4), of the transition element between positive energy free electron states gives the Møller scattering of two electrons, when account is taken of the Pauli principle.

The exclusion principle for interacting charges is handled in exactly the same way as for noninteracting charges (I). For example, for two charges it requires only that one calculate  $K(3, 4; 1, 2) - K(4, 3; 1, 2)$  to get the net

amplitude for arrival of charges at 3 and 4. It is disregarded in intermediate states. The interference effects for scattering of electrons by positrons discussed by Bhabha will be seen to result directly in this formulation. The formulas are interpreted to apply to positrons in the manner discussed in I.

As our primary concern will be for processes in which the quanta are virtual we shall not include here the detailed analysis of processes involving real quanta in initial or final state, and shall content ourselves by only stating the rules applying to them.<sup>9</sup> The result of the analysis is, as expected, that they can be included by the same line of reasoning as is used in discussing the virtual processes, provided the quantities are normalized in the usual manner to represent single quanta. For example, the amplitude that an electron in going from 1 to 2 absorbs a quantum whose vector potential, suitably normalized, is  $c_\mu \exp(-ik \cdot x) = C_\mu(x)$  is just the expression (I, Eq. (13)) for scattering in a potential with  $\mathbf{A}$  (3) replaced by  $\mathbf{C}$  (3). Each quantum interacts only once (either in emission or in absorption), terms like (I, Eq. (14)) occur only when there is more than one quantum involved. The Bose statistics of the quanta can, in all cases, be disregarded in intermediate states. The only effect of the statistics is to change the weight of initial or final states. If there are among quanta, in the initial state, some which are identical then the weight of the state is  $(1/n!)$  of what it would be if these quanta were considered as different (similarly for the final state).

### 3 THE SELF-ENERGY PROBLEM

Having a term representing the mutual interaction of a pair of charges, we must include similar terms to represent the interaction of a charge with itself. For under some circumstances what appears to be two distinct electrons

---

<sup>9</sup>Although in the expressions stemming from (4) the quanta are virtual, this is not actually a theoretical limitation. One way to deduce the correct rules for real quanta from (4) is to note that in a closed system all quanta can be considered as virtual (i.e., they have a known source and are eventually absorbed) so that in such a system the present description is complete and equivalent to the conventional one. In particular, the relation of the Einstein  $A$  and  $B$  coefficients can be deduced. A more practical direct deduction of the expressions for real quanta will be given in the subsequent paper. It might be noted that (4) can be rewritten as describing the action on  $a$ ,  $K^{(1)}(3,1) = i \int K_+(3,5) \times A(5)K_+(5,1)d\tau_5$  of the potential  $A_\mu(5) = e^2 \int K_+(4,6)\delta_+(s_{56}^2)\gamma_\mu \times K_+(6,2)d\tau_6$  arising from Maxwell's equations  $-\square^2 A_\mu = 4\pi j_\mu$  from a "current"  $j_\mu(6) = e^2 K_+(4,6)\gamma_\mu K_+(6,2)$  produced by particle  $b$  in going from 2 to 4. This is virtue of the fact that  $\delta_+$  satisfies

$$-\square_2^2 \delta_+(s_{21}^2) = 4\pi\delta(2,1). \quad (5)$$

may, according to I, be viewed also as a single electron (namely in case one electron was created in a pair with a positron destined to annihilate the other electron). Thus to the interaction between such electrons must correspond the possibility of the action of an electron on itself.<sup>10</sup>

This interaction is the heart of the self energy problem. Consider to first order in  $e^2$  the action of an electron on itself in an otherwise force free region. The amplitude  $K(2, 1)$  for a single particle to get from 1 to 2 differs from  $K_+(2, 1)$  to first order in  $e^2$  by a term

$$K^{(1)}(2, 1) = -ie^2 \int \int K_+(2, 4)\gamma_\mu K_+(4, 3)\gamma_\mu \\ \times K_+(3, 1)d\tau_3 d\tau_4 \delta_+(s_{43}^2). \quad (6)$$

It arises because the electron instead of going from 1 directly to 2, may go (Fig. 2) first to 3, ( $K_+(3, 1)$ ), emit a quantum ( $\gamma_\mu$ ), proceed to 4, ( $K_+(4, 3)$ ), absorb it ( $\gamma_\mu$ ), and finally arrive at 2 ( $K_+(2, 4)$ ). The quantum must go from 3 to 4 ( $\delta_+(s_{43}^2)$ ).

This is related to the self-energy of a free electron in the following manner. Suppose initially, time  $t_1$ , we have an electron in state  $f(1)$  which we imagine to be a positive energy solution of Dirac's equation for a free particle. After a long time  $t_2 - t_1$  the perturbation will alter the wave function, which can then be looked upon as a superposition of free particle solutions (actually it only contains  $f$ ). The amplitude that  $g(2)$  is contained is calculated as in (I, Eq. (21)). The diagonal element ( $g = f$ ) is therefore

$$\int \int \tilde{f}(2)\beta K^{(1)}(2, 1)\beta f(1)d^3x_1 d^3x_2. \quad (7)$$

The time interval  $T = t_2 - t_1$  (and the spatial volume  $V$  over which one integrates) must be taken very large, for the expressions are only approximate (analogous to the situation for two interacting charges).<sup>11</sup> This is because, for example, we are dealing incorrectly with quanta emitted just before  $t_2$  which would normally be reabsorbed at times after  $t_2$ .

If  $K^{(1)}(2, 1)$  from (6) is actually substituted into (7) the surface integrals can be performed as was done in obtaining I, Eq. (22) resulting in

$$-ie^2 \int \int \tilde{f}(4)\gamma_\mu K_+(4, 3)\gamma_\mu f(3)\delta_+(s_{43}^2)d\tau_3 d\tau_4. \quad (8)$$

---

<sup>10</sup>These considerations make it appear unlikely that the contention of J. A. Wheeler and R. P. Feynman, Rev. Mod. Phys. **17**, 157 (1945), that electrons do not act on themselves, will be a successful concept in quantum electrodynamics.

<sup>11</sup>This is discussed in reference 5 in which it is pointed out that the concept of a wave function loses accuracy if there are delayed self-actions.

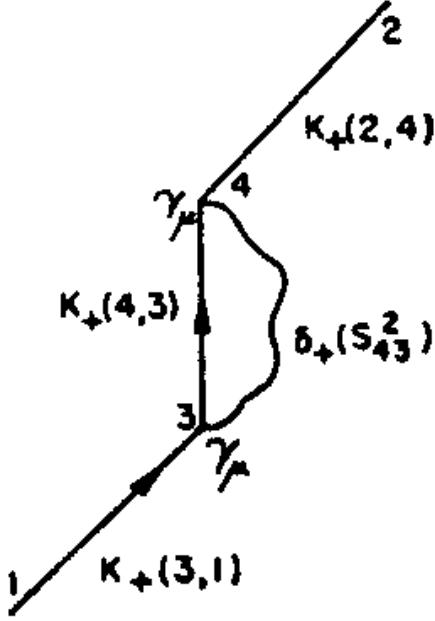


Figure 2: Interaction of an electron with itself, Eq. (6).

Putting for  $f(1)$  the plane wave  $u \exp(-ip \cdot x_1)$  where  $p_\mu$  is the energy ( $p_4$ ) and momentum of the electron ( $\mathbf{p}^2 = m^2$ ), and  $u$  is a constant 4-index symbol, (8) becomes

$$-ie^2 \int \int (\tilde{u} \gamma_\mu K_+(4,3) \gamma_\mu u) \times \exp(ip \cdot (x_4 - x_2)) \delta_+(s^2_{43}) d\tau_3 d\tau_4,$$

the integrals extending over the volume  $V$  and time interval  $T$ . Since  $K_+(4,3)$  depends only on the difference of the coordinates of 4 and 3,  $x_{43\mu}$ , the integral on 4 gives a result (except near the surfaces of the region) independent of 3. When integrated on 3, therefore, the result is of order  $VT$ . The effect is proportional to  $V$ , for the wave functions have been normalized to unit volume. If normalized to volume  $V$ , the result would simply be proportional to  $T$ . This is expected, for if the effect were equivalent to a change in energy  $\Delta E$ , the amplitude for arrival in  $f$  at  $t_2$  is altered by a factor  $\exp(-i\Delta E(t_2 - t_1))$ , or to first order by the difference  $-i(\Delta E)T$ .

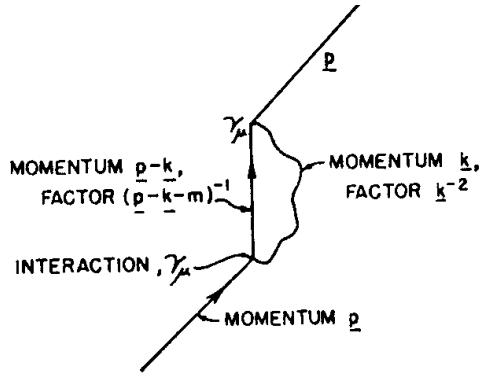


Figure 3: Interaction of an electron with itself. Momentum space, Eq. (11).

Hence, we have

$$\Delta E = e^2 \int (\tilde{u} \gamma_\mu K_+(4, 3) \gamma_\mu u) \exp(ip \cdot x_{43}) \delta_+(s_{43}^2) d\tau_4, \quad (9)$$

integrated over all space-time  $d\tau_4$ . This expression will be simplified presently. In interpreting (9) we have tacitly assumed that the wave functions are normalized so that  $(u^* u) = (\tilde{u} \gamma_4 u) = 1$ . The equation may therefore be made independent of the normalization by writing the left side as  $(\Delta E)(\tilde{u} \gamma_4 u)$ , or since  $(\tilde{u} \gamma_4 u) = (E/m)(\tilde{u} u)$  and  $m \Delta m = E \Delta E$ , as  $\Delta m(\tilde{u} u)$  where  $\Delta m$  is an equivalent change in mass of the electron. In this form invariance is obvious.

One can likewise obtain an expression for the energy shift for an electron in a hydrogen atom. Simply replace  $K_+$  in (8), by  $K_+^{(V)}$ , the exact kernel for an electron in the potential,  $\mathbf{V} = \beta e^2/r$ , of the atom, and  $f$  by a wave function (of space and time) for an atomic state. In general the  $\Delta E$  which results is not real. The imaginary part is negative and in  $\exp(-i\Delta ET)$  produces an exponentially decreasing amplitude with time. This is because we are asking for the amplitude that an atom initially with no photon in the field, will still appear after time  $T$  with no photon. If the atom is in a state which can radiate, this amplitude must decay with time. The imaginary part of  $\Delta E$  when calculated does indeed give the correct rate of radiation from atomic states. It is zero for the ground state and for a free electron.

In the non-relativistic region the expression for  $\Delta E$  can be worked out as has been done by Bethe.<sup>12</sup> In the relativistic region (points 4 and 3 as

---

<sup>12</sup>H. A. Bethe, Phys. Rev. **72**, 339 (1947).

close together as a Compton wave-length) the  $K_+^{(V)}$  which should appear in (8) can be replaced to first order in  $\mathbf{V}$  by  $K_+$  plus  $K_+^{(1)}(2, 1)$  given in I, Eq. (13). The problem is then very similar to the radiationless scattering problem discussed below.

#### 4 EXPRESSION IN MOMENTUM AND ENERGY SPACE

The evaluation of (9), as well as all the other more complicated expressions arising in these problems, is very much simplified by working in the momentum and energy variables, rather than space and time. For this we shall need the Fourier Transform of  $\delta_+(s_{21}^2)$  which is

$$-\delta_+(s_{21}^2) = \pi^{-1} \int \exp(-ik \cdot x_{21}) \mathbf{k}^{-2} d^4 k, \quad (10)$$

which can be obtained from (3) and (5) or from I, Eq. (32) noting that  $I_+(2, 1)$  for  $m^2 = 0$  is  $\delta_+(s_{21}^2)$  from I, Eq. (34). The  $\mathbf{k}^{-2}$  means  $(k \cdot k)^{-1}$

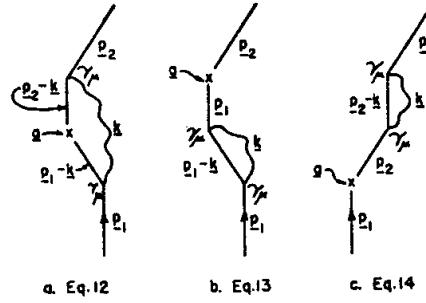


Figure 4: Radiative correction to scattering, momentum space.

or more precisely the limit as  $\delta \rightarrow 0$  of  $(k \cdot k + i\delta)^{-1}$ . Further  $d^4 k$  means  $(2\pi)^{-2} dk_1 dk_2 dk_3 dk_4$ . If we imagine that quanta are particles of zero mass, then we can make the general rule that all poles are to be resolved by considering the masses of the particles and quanta to have infinitesimal negative imaginary parts.

Using these results we see that the self-energy (9) is the matrix element between  $u$  and  $u'$  of the matrix

$$(e^2/\pi i) \int \gamma_\mu (\mathbf{p} - \mathbf{k} - m)^{-1} \gamma_\mu \mathbf{k}^{-2} d^4 k, \quad (11)$$

where we have used the expression (I, Eq. (31)) for the Fourier transform of  $K_+$ . This form for the self-energy is easier to work with than is (9).

The equation can be understood by imagining (Fig. 3) that the electron of momentum  $\mathbf{p}$  emits ( $\gamma_\mu$ ) a quantum of momentum  $\mathbf{k}$ , and makes its way now with momentum  $\mathbf{p} - \mathbf{k}$  to the next event (factor  $(\mathbf{p} - \mathbf{k} - m)^{-1}$ ) which is to absorb the quantum (another  $\gamma_\mu$ ). The amplitude of propagation of quanta is  $\mathbf{k}^{-2}$ . (There is a factor  $e^2/\pi i$  for each virtual quantum). One integrates over all quanta. The reason an electron of momentum  $\mathbf{p}$  propagates as  $1/(\mathbf{p} - m)$  is that this operator is the reciprocal of the Dirac equation operator, and we are simply solving this equation. Likewise light goes as  $1/\mathbf{k}^2$ , for this is the reciprocal D'Alembertian operator of the wave equation of light. The first  $\gamma_\mu$ , represents the current which generates the vector potential, while the second is the velocity operator by which this potential is multiplied in the Dirac equation when an external field acts on an electron.

Using the same line of reasoning, other problems may be set up directly in momentum space. For example, consider the scattering in a potential  $\mathbf{A} = A_\mu \gamma_\mu$  varying in space and time as  $\mathbf{a} \exp(-iq \cdot x)$ . An electron initially in state of momentum  $\mathbf{p}_1 = p_{1\mu} \gamma_\mu$  will be deflected to state  $\mathbf{p}_2$  where  $\mathbf{p}_2 = \mathbf{p}_1 + \mathbf{q}$ . The zero-order answer is simply the matrix element of  $\mathbf{a}$  between states 1 and 2. We next ask for the first order (in  $e^2$ ) radiative correction due to virtual radiation of one quantum. There are several ways this can happen. First for the case illustrated in Fig. 4(a), find the matrix:

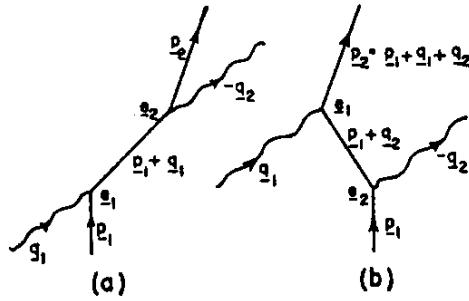


Figure 5: Compton scattering, Eq. (15).

$$(e^2/\pi i) \int \gamma_\mu (\mathbf{p}_2 - \mathbf{k} - m)^{-1} \mathbf{a} (\mathbf{p}_1 - \mathbf{k} - m)^{-1} \gamma_\mu \mathbf{k}^{-2} d^4 k. \quad (12)$$

For in this case, first<sup>13</sup> a quantum of momentum  $\mathbf{k}$  is emitted ( $\gamma_\mu$ ), the electron then having momentum  $\mathbf{p}_1 - \mathbf{k}$  and hence propagating with factor  $(\mathbf{p}_1 - \mathbf{k} - m)^{-1}$ . Next it is scattered by the potential (matrix  $\mathbf{a}$ ) receiving additional momentum  $\mathbf{q}$ , propagating on then (factor  $(\mathbf{p}_2 - \mathbf{k} - m)^{-1}$ ) with the new momentum until the quantum is reabsorbed ( $\gamma_\mu$ ). The quantum propagates from emission to absorption ( $\mathbf{k}^{-2}$ ) and we integrate over all quanta ( $d^4k$ ), and sum on polarization  $\mu$ . When this is integrated on  $k_4$ , the result can be shown to be exactly equal to the expressions (16) and (17) given in  $B$  for the same process, the various terms coming from residues of the poles of the integrand (12).

Or again if the quantum is both emitted and reabsorbed before the scattering takes place one finds (Fig. 4(b))

$$(e^2/\pi i) \int \mathbf{a}(\mathbf{p}_1 - m)^{-1} \gamma_\mu (\mathbf{p}_1 - \mathbf{k} - m)^{-1} \gamma_\mu \mathbf{k}^{-2} d^4k, \quad (13)$$

or if both emission and absorption occur after the scattering, (Fig. 4(c))

$$(e^2/\pi i) \int \gamma_\mu (\mathbf{p}_2 - \mathbf{k} - m)^{-1} \gamma_\mu (\mathbf{p}_2 - m)^{-1} \mathbf{a} \mathbf{k}^{-2} d^4k. \quad (14)$$

These terms are discussed in detail below.

We have now achieved our simplification of the form of writing matrix elements arising from virtual processes. Processes in which a number of real quanta is given initially and finally offer no problem (assuming correct normalization). For example, consider the Compton effect (Fig. 5(a)) in which an electron in state  $\mathbf{p}_1$  absorbs a quantum of momentum  $\mathbf{q}_1$ , polarization vector  $e_1\mu$  so that its interaction is  $e_1\mu\gamma_\mu = \mathbf{e}_1$ , and emits a second quantum of momentum  $-\mathbf{q}_2$  polarization  $\mathbf{e}_2$  to arrive in final state of momentum  $\mathbf{p}_2$ . The matrix for this process is  $\mathbf{e}_2(\mathbf{p}_1 + \mathbf{q}_1 - m)^{-1}\mathbf{e}_1$ . The total matrix for the Compton effect is, then,

$$\mathbf{e}_2(\mathbf{p}_1 + \mathbf{q}_1 - m)^{-1}\mathbf{e}_1 + \mathbf{e}_1(\mathbf{p}_1 + \mathbf{q}_2 - m)^{-1}\mathbf{e}_3, \quad (15)$$

the second term arising because the emission of  $\mathbf{e}_2$ , may also precede the absorption of  $\mathbf{e}_1$  (Fig. 5(b)). One takes matrix elements of this between initial and final electron states ( $\mathbf{p}_1 + \mathbf{q}_1 = \mathbf{p}_2 - \mathbf{q}_2$ ), to obtain the Klein Nishina formula. Pair annihilation with emission of two quanta, etc., are given by the same matrix, positron states being those with negative time component of  $\mathbf{p}$ . Whether quanta are absorbed or emitted depends on whether the time component of  $\mathbf{q}$  is positive or negative.

---

<sup>13</sup>First, next, etc., here refer not to the order in true time but to the succession of events along the trajectory of the electron. That is, more precisely, to the order of appearance of the matrices in the expressions.

## 5 THE CONVERGENCE OF PROCESSES WITH VIRTUAL QUANTA

These expressions are, as has been indicated, no more than a re-expression of conventional quantum electrodynamics. As a consequence, many of them are meaningless. For example, the self-energy expression (9) or (11) gives an infinite result when evaluated. The infinity arises, apparently, from the coincidence of the  $\delta$ -function singularities in  $K_+(4,3)$  and  $\delta_+(s_{43}^2)$ . Only at this point is it necessary to make a real departure from conventional electrodynamics, a departure other than simply rewriting expressions in a simpler form.

We desire to make a modification of quantum electrodynamics analogous to the modification of classical electrodynamics described in a previous article, *A*. There the  $\delta(s_{12}^2)$  appearing in the action of interaction was replaced by  $f(s_{12}^2)$  where  $f(x)$  is a function of small width and great height.

The obvious corresponding modification in the quantum theory is to replace the  $\delta_+(s^2)$  appearing in the quantum mechanical interaction by a new function  $f_-(s^2)$ . We can postulate that if the Fourier transform of the classical  $f(s_{12}^2)$  is the integral over all  $\mathbf{k}$  of  $F(\mathbf{k}^2)\exp(-ik \cdot x_{12})d^4k$ , then the Fourier transform of  $f_-(s^2)$  is the same integral taken over only positive frequencies  $k_4$  for  $t_2 > t_1$  and over only negative ones for  $t_2 < t_1$  in analogy to the relation of  $\delta_+(s^2)$  to  $\delta(s^2)$ . The function  $f(s^2) = f(x \cdot x)$  can be written <sup>14</sup> as

$$f(x \cdot x) = (2\pi)^{-2} \int_{k_4=0}^{\infty} \int \sin(k_4|x_4|) \times \cos(\mathbf{K} \cdot \mathbf{x}) dk_4 d^3 \mathbf{K} g(k \cdot k),$$

where  $g(k \cdot k)$  is  $k_4^{-1}$  times the density of oscillators and may be expressed for positive  $k_4$  as (*A*, Eq. (16))

$$g(\mathbf{k}^2) = \int_0^{\infty} (\delta(\mathbf{k}^2) - \delta(\mathbf{k}^2 - \lambda^2)) G(\lambda) d\lambda,$$

where  $\int_0^{\infty} G(\lambda) d\lambda = 1$  and  $G$  involves values of  $\lambda$  large compared to  $m$ . This simply means that the amplitude for propagation of quanta of momentum

---

<sup>14</sup>This relation is given incorrectly in *A*, equation just preceding 16.

$\mathbf{k}$  is

$$-F_+(\mathbf{k}^3) = \pi^{-1} \int_0^\infty (\mathbf{k}^{-2} - (\mathbf{k}^2 - \lambda^2)^{-1}) G(\lambda) d\lambda,$$

rather than  $\mathbf{k}^{-2}$ . That is, writing  $F_+(\mathbf{k}^2) = -\pi^{-1}\mathbf{k}^{-2}C(\mathbf{k}^2)$ ,

$$-f_+(s_{12}^2) = \pi^{-1} \int \exp(-ik \cdot x_{12}) \mathbf{k}^{-2} C(\mathbf{k}^2) d^4k. \quad (16)$$

Every integral over an intermediate quantum which previously involved a factor  $d^4k/\mathbf{k}^2$  is now supplied with a convergence factor  $C(\mathbf{k}^2)$  where

$$C(\mathbf{k}^2) = \int_0^\infty -\lambda^2 (\mathbf{k}^2 - \lambda^2)^{-1} G(\lambda) d\lambda. \quad (17)$$

The poles are defined by replacing  $\mathbf{k}^2$  by  $\mathbf{k}^2 + i\delta$  in the limit  $\delta \rightarrow 0$ . That is  $\lambda^2$  may be assumed to have an infinitesimal negative imaginary part.

The function  $f_+(s_{12}^2)$  may still have a discontinuity in value on the light cone. This is of no influence for the Dirac electron. For a particle satisfying the Klein Gordon equation, however, the interaction involves gradients of the potential which reinstates the  $\delta$  function if  $f$  has discontinuities. The condition that  $f$  is to have no discontinuity in value on the light cone implies  $\mathbf{k}^2 C(\mathbf{k}^2)$  approaches zero as  $\mathbf{k}^2$  approaches infinity. In terms of  $G(\lambda)$  the condition is

$$\int_0^\infty \lambda^2 G(\lambda) d\lambda = 0. \quad (18)$$

This condition will also be used in discussing the convergence of vacuum polarization integrals.

The expression for the self-energy matrix is now

$$(e^2/\pi i) \int \gamma_\mu (\mathbf{p} - \mathbf{k} - m)^{-1} \gamma_\mu \mathbf{k}^{-2} d^4k C(\mathbf{k}^2), \quad (19)$$

which, since  $C(\mathbf{k}^2)$  falls off at least as rapidly as  $1/\mathbf{k}^2$ , converges. For practical purposes we shall suppose hereafter that  $C(\mathbf{k}^2)$  is simply  $-\lambda^2/(\mathbf{k}^2 - \lambda^2)$  implying that some average (with weight  $G(\lambda)d\lambda$ ) over values of  $\lambda$  may be taken afterwards. Since in all processes the quantum momentum will be contained in at least one extra factor of the form  $(\mathbf{p} - \mathbf{k} - m)^{-1}$  representing propagation of an electron while that quantum is in the field, we can expect all such integrals with their convergence factors to converge and that the result of all such processes will now be finite and definite (excepting the

processes with closed loops, discussed below, in which the diverging integrals are over the momenta of the electrons rather than the quanta).

The integral of (19) with  $C(\mathbf{k}^2) = -\lambda^2(\mathbf{k}^2 - \lambda^2)^{-1}$  noting that  $\mathbf{p}^2 = m^2$ ,  $\lambda \gg m$  and dropping terms of order  $m/\lambda$  is (see Appendix A)

$$(e^2/2\pi)[4m(\ln(\lambda/m) + \frac{1}{2}) - \mathbf{p}(\ln(\lambda/m) + 5/4)]. \quad (20)$$

When applied to a state of an electron of momentum  $\mathbf{p}$  satisfying  $\mathbf{p}u = mu$ , it gives for the change in mass (as in B, Eq. (9))

$$\Delta m = m(e^2/2\pi)(3\ln(\lambda/m) + \frac{3}{4}). \quad (21)$$

## 6 RADIATIVE CORRECTIONS TO SCATTERING

We can now complete the discussion of the radiative corrections to scattering. In the integrals we include the convergence factor  $C(\mathbf{k}^2)$ , so that they converge for large  $\mathbf{k}$ . Integral (12) is also not convergent because of the well-known infrared catastrophe. For this reason we calculate (as discussed in B) the value of the integral assuming the photons to have a small mass  $\lambda_{\min} \ll m \ll \lambda$ . The integral (12) becomes

$$(e^2/\pi i) \int \gamma_\mu (\mathbf{p}_2 - \mathbf{k} - m)^{-1} \mathbf{a} (\mathbf{p}_1 - \mathbf{k} - m)^{-1} \\ \times \gamma_\mu (\mathbf{k}^2 - \gamma_{\min}^2)^{-1} d^4 k C(\mathbf{k}^2 - \lambda_{\min}^2),$$

which when integrated (see Appendix B) gives  $(e^2/2\pi)$  times

$$\left[ 2 \left( \ln \frac{m}{\lambda_{\min}} - 1 \right) \left( 1 - \frac{2\theta}{\tan 2\theta} \right) + \theta \tan \theta \right. \\ \left. + \frac{4}{\tan 2\theta} \int_0^\theta \alpha \tan \alpha d\alpha \right] \mathbf{a} + \frac{1}{4m} (\mathbf{q}\mathbf{a} - \mathbf{a}\mathbf{q}) \frac{2\theta}{\sin 2\theta} + r\mathbf{a}, \quad (22)$$

where  $(\mathbf{q}^2)^{1/2} = 2m$  and we have assumed the matrix to operate between states of momentum  $\mathbf{p}_1$  and  $\mathbf{p}_2 = \mathbf{p}_1 + \mathbf{q}$  and have neglected terms of order  $\lambda_{\min}/m, m/\lambda$ , and  $\mathbf{q}^2/\lambda^2$ . Here the only dependence on the convergence factor is in the term  $r\mathbf{a}$ , where

$$\mathbf{r} = \ln(\lambda/m) + 9/4 - 2\ln(m/\lambda_{\min}). \quad (23)$$

As we shall see in a moment, the other terms (13), (14) give contributions which just cancel the  $\mathbf{r}\mathbf{a}$  term. The remaining terms give for small  $\mathbf{q}$ ,

$$(e^2/4\pi) \left( \frac{1}{2m}(\mathbf{q}\mathbf{a} - \mathbf{a}\mathbf{q}) + \frac{4\mathbf{q}^2}{3m^2}\mathbf{a} \left( \ln \frac{m}{\lambda_{\min}} - \frac{3}{8} \right) \right), \quad (24)$$

which shows the change in magnetic moment and the Lamb shift as interpreted in more detail in B.<sup>15</sup>

We must now study the remaining terms (13) and (14). The integral on  $\mathbf{k}$  in (13) can be performed (after multiplication by  $C(\mathbf{k}^2)$ ) since it involves nothing but the integral (19) for the self-energy and the result is allowed to operate on the initial state  $u_1$ , (so that  $\mathbf{p}_1 u_1 = m u_1$ ). Hence the factor following  $\mathbf{a}(\mathbf{p}_1 - m)^{-1}$  will be just  $\Delta m$ . But, if one now tries to expand  $1/(\mathbf{p}_1 - m) = (\mathbf{p}_1 + m)/(\mathbf{p}_1^2 - m^2)$  one obtains an infinite result, since  $\mathbf{p}_1^2 = m^2$ . This is, however, just what is expected physically. For the quantum can be emitted and absorbed at any time previous to the scattering. Such a process has the effect of a change in mass of the electron in the state 1. It therefore changes the energy by  $\Delta E$  and the amplitude to first order in  $\Delta E$  by  $-i\Delta E \cdot t$  where  $t$  is the time it is acting, which is infinite. That is, the major effect of this term would be canceled by the effect of change of mass  $\Delta m$ .

The situation can be analyzed in the following manner. We suppose that the electron approaching the scattering potential  $\mathbf{a}$  has not been free for an infinite time, but at some time far past suffered a scattering by a potential  $\mathbf{b}$ . If we limit our discussion to the effects of  $\Delta m$  and of the virtual radiation of one quantum between two such scatterings each of the effects will be finite, though large, and their difference is determinate. The propagation from  $\mathbf{b}$  to  $\mathbf{a}$  is represented by a matrix

$$\mathbf{a}(\mathbf{p}' - m)^{-1}\mathbf{b}, \quad (25)$$

---

<sup>15</sup>That the result given in B in Eq. (19) was in error was repeatedly pointed out to the author, in private communication, by V. F. Weisskopf and J. B. French, as their calculation, completed simultaneously with the author's early in 1948, gave a different result. French has finally shown that although the expression for the radiationless scattering B, Eq. (18) or (24) above is correct, it was incorrectly joined into Bethe's non-relativistic result. He shows that the relation  $\ln 2k_{\max} - 1 = \ln \lambda_{\min}$  used by the author should have been  $\ln 2k_{\max} - 5/6 = \ln \lambda_{\min}$ . This results in adding a term  $-(1/6)$  to the logarithm in B, Eq. (19) so that the result now agrees with that of J. B. French and V. F. Weisskopf, Phys. Rev. **75**, 1240 (1949) and N. H. Kroll and W. E. Lamb, Phys. Rev. **75**, 388 (1949). The author feels unhappily responsible for the very considerable delay in the publication of French's result occasioned by this error. This footnote is appropriately numbered.

in which one is to integrate possibly over  $\mathbf{p}'$  (depending on details of the situation). (If the time is long between  $\mathbf{b}$  and  $\mathbf{a}$ , the energy is very nearly determined so that  $\mathbf{p}'^2$  is very nearly  $m^2$ .)

We shall compare the effect on the matrix (25) of the virtual quanta and of the change of mass  $\Delta m$ . The effect of a virtual quantum is

$$(e^2/\pi i) \int \mathbf{a}(\mathbf{p}' - m)^{-1} \gamma_\mu (\mathbf{p}' - \mathbf{k} - m)^{-1} \times \gamma_\mu (\mathbf{p}' - m)^{-1} \mathbf{b} \mathbf{k}^{-2} d^4 k C(\mathbf{k}^2), \quad (26)$$

while that of a change of mass can be written

$$\mathbf{a}(\mathbf{p}' - m)^{-1} \Delta m (\mathbf{p}' - m)^{-1} \mathbf{b}, \quad (27)$$

and we are interested in the difference (26)-(27). A simple and direct method of making this comparison is just to evaluate the integral on  $\mathbf{k}$  in (26) and subtract from the result the expression (27) where  $\Delta m$  is given in (21). The remainder can be expressed as a multiple  $-r(\mathbf{p}'^2)$  of the unperturbed amplitude (25);

$$-r(\mathbf{p}'^2) \mathbf{a}(\mathbf{p}' - m)^{-1} \mathbf{b}. \quad (28)$$

This has the same result (to this order) as replacing the potentials  $\mathbf{a}$  and  $\mathbf{b}$  in (25) by  $(1 - \frac{1}{2}r(\mathbf{p}'^2))\mathbf{a}$  and  $(1 - \frac{1}{2}\mathbf{r}(\mathbf{p}'^2))\mathbf{b}$ . In the limit, then, as  $\mathbf{p}'^2 \rightarrow m^2$  the net effect on the scattering is  $-\frac{1}{2}\mathbf{r}\mathbf{a}$  where  $\mathbf{r}$ , the limit of  $\mathbf{r}(\mathbf{p}'^2)$  as  $\mathbf{p}'^2 \rightarrow m^2$  (assuming the integrals have an infrared cut-off), turns out to be just equal to that given in (23). An equal term  $-\frac{1}{2}\mathbf{r}\mathbf{a}$  arises from virtual transitions after the scattering (14) so that the entire  $\mathbf{r}\mathbf{a}$  term in (22) is canceled.

The reason that  $\mathbf{r}$  is just the value of (12) when  $\mathbf{q}^2 = 0$  can also be seen without a direct calculation as follows: Let us call  $\mathbf{p}$  the vector of length  $m$  in the direction of  $\mathbf{p}'$  so that if  $\mathbf{p}'^2 = m(1 + \epsilon)^2$  we have  $\mathbf{p}' = (1 + \epsilon)\mathbf{p}$  and we take  $\epsilon$  as very small, being of order  $T^{-1}$  where  $T$  is the time between the scatterings  $\mathbf{b}$  and  $\mathbf{a}$ . Since  $(\mathbf{p}' - m)^{-1} = (\mathbf{p}' + m)/(\mathbf{p}'^2 - m^2) \approx (\mathbf{p}' + m)/2m^2\epsilon$ , the quantity (25) is of order  $\epsilon^{-1}$  or  $T$ . We shall compute corrections to it only to its own order ( $\epsilon^{-1}$ ) in the limit  $\epsilon \rightarrow 0$ . The term (27) can be written approximately<sup>16</sup> as

$$(e^2/\pi i) \int \mathbf{a}(\mathbf{p}' - m)^{-1} \gamma_\mu (\mathbf{p}' - \mathbf{k} - m)^{-1} \times \gamma_\mu (\mathbf{p}' - m)^{-1} \mathbf{b} \mathbf{k}^{-2} d^4 k C(\mathbf{k}^2),$$

---

<sup>16</sup>The expression is not exact because the substitution of  $\Delta m$  by the integral in (19) is valid only if  $\mathbf{p}$  operates on a state such that  $\mathbf{p}$  can be replaced by  $m$ . The error, however, is of order  $\mathbf{a}(\mathbf{p}' - m)^{-1}(\mathbf{p}' - m)(\mathbf{p}' - m)^{-1}\mathbf{b}$  which is  $\mathbf{a}((1 + \epsilon)\mathbf{p} + m)(\mathbf{p}' - m) \times ((1 + \epsilon)\mathbf{p} + m)\mathbf{p}(2\epsilon + \epsilon^2)^{-2}m^{-4}$ . But since  $\mathbf{p}'^2 = m^2$  we have  $\mathbf{p}(\mathbf{p}' - m) = -m(\mathbf{p}' - m) = (\mathbf{p}' - m)\mathbf{p}$  so the net result is approximately  $\mathbf{a}(\mathbf{p}' - m)\mathbf{b}/4m^2$  and is not of order  $1/\epsilon$  but smaller, so that its effect drops out in the limit.

using the expression (19) for  $\Delta m$ . The net of the two effects is therefore approximately<sup>17</sup>

$$-(e^2/\pi i) \int \mathbf{a}(\mathbf{p}' - m)^{-1} \gamma_\mu (\mathbf{p} - \mathbf{k} - m)^{-1} \epsilon \mathbf{p} (\mathbf{p} - \mathbf{k} - m)^{-1} \\ \times \gamma_\mu (\mathbf{p}' - m)^{-1} \mathbf{b} \mathbf{k}^{-2} d^4 k C(\mathbf{k}^2),$$

a term now of order  $1/\epsilon$  (since  $(\mathbf{p}' - m)^{-1} \approx (\mathbf{p} + m) \times (2m^2\epsilon)^{-1}$ ) and therefore the one desired in the limit. Comparison to (28) gives for  $\mathbf{r}$  the expression

$$(\mathbf{p}_1 + m/2m) \int \gamma_\mu (\mathbf{p}_1 - \mathbf{k} - m)^{-1} (\mathbf{p}_1 m^{-1}) (\mathbf{p}_1 - \mathbf{k} - m)^{-1} \\ \times \gamma_\mu \mathbf{k}^{-2} d^4 k C(\mathbf{k}^2). \quad (29)$$

The integral can be immediately evaluated, since it is the same as the integral (12), but with  $\mathbf{q} = 0$ , for  $\mathbf{a}$  replaced by  $\mathbf{p}_1/m$ . The result is therefore  $\mathbf{r} \cdot (\mathbf{p}_1/m)$  which when acting on the state  $u_1$  is just  $\mathbf{r}$ , as  $\mathbf{p}_1 u_1 = mu_1$ . For the same reason the term  $(\mathbf{p}_1 + m)/2m$  in (29) is effectively 1 and we are left with  $-\mathbf{r}$  of (23).<sup>18</sup>

In more complex problems starting with a free electron the same type of term arises from the effects of a virtual emission and absorption both previous to the other processes. They, therefore, simply lead to the same factor  $\mathbf{r}$  so that the expression (23) may be used directly and these renormalization integrals need not be computed afresh for each problem.

In this problem of the radiative corrections to scattering the net result is insensitive to the cut-off. This means, of course, that by a simple rearrangement of terms previous to the integration we could have avoided the use of the convergence factors completely (see for example Lewis<sup>19</sup>). The problem was solved in the manner here in order to illustrate how the use of

---

<sup>17</sup>We have used, to first order, the general expansion (valid for any operators  $A, B$ )

$$(A + B 0^{-1}) = A^{-1} - A^{-1} B A^{-1} + A^{-1} B A^{-1} B A^{-1} - \dots$$

with  $A = \mathbf{p} - \mathbf{k} - m$  and  $B = \mathbf{p}' - \mathbf{p} = \epsilon \mathbf{p}$  to expand the difference of  $(\mathbf{p}' - \mathbf{k} - m)^{-1}$  and  $(\mathbf{p} - \mathbf{k} - m)^{-1}$ .

<sup>18</sup>The renormalization terms appearing  $B$ , Eqs. (14), (15) when translated directly into the present notation do not give twice (29) but give this expression with the central  $\mathbf{p}_1 m^{-1}$  factor replaced by  $m\gamma_4/E_1$  where  $E_1 = p_{1\mu}$ , for  $\mu = 4$ . When integrated it therefore gives  $\mathbf{r} \mathbf{a}((\mathbf{p}_1 + m)/2m)(m\gamma_4/E_1)$  or  $\mathbf{r} \mathbf{a} - \mathbf{r} \mathbf{a}(m\gamma_4/E_1)(\mathbf{p}_1 - m)/2m$ . (Since  $\mathbf{p}_1 \gamma_4 + \gamma_4 \mathbf{p}_1 = 2E_1$ ) which gives just  $r_a$ , since  $\mathbf{p}_1 u_1 = mu_1$ .

<sup>19</sup>H.W. Lewis, Phys. Rev. **73**, 173 (1948).

such convergence factors, even when they are actually unnecessary, may facilitate analysis somewhat by removing the effort and ambiguities that may be involved in trying to rearrange the otherwise divergent terms.

The replacement of  $\delta_+$  by  $f_+$  given in (16), (17) is not determined by the analogy with the classical problem. In the classical limit only the real part of  $\delta_+$  (i.e., just  $\delta$ ) is easy to interpret. But by what should the imaginary part,  $1/(\pi s^2)$ , of  $\delta_+$  be replaced? The choice we have made here (in denning, as we have, the location of the poles of (17)) is arbitrary and almost certainly incorrect. If the radiation resistance is calculated for an atom, as the imaginary part of (8), the result depends slightly on the function  $f_+$ . On the other hand the light radiated at very large distances from a source is independent of  $f_+$ . The total energy absorbed by distant absorbers will not check with the energy loss of the source. We are in a situation analogous to that in the classical theory if the entire  $f$  function is made to contain only retarded contributions (see A, Appendix). One desires instead the analogue of  $\langle F \rangle_{\text{ret}}$  of A. This problem is being studied.

One can say therefore, that this attempt to find a consistent modification of quantum electrodynamics is incomplete (see also the question of closed loops, below). For it could turn out that any correct form of  $f_+$  which will guarantee energy conservation may at the same time not be able to make the self-energy integral finite. The desire to make the methods of simplifying the calculation of quantum electrodynamic processes more widely available has prompted this publication before an analysis of the correct form for  $f_+$  is complete. One might try to take the position that, since the energy discrepancies discussed vanish in the limit  $\lambda \rightarrow \infty$ , the correct physics might be considered to be that obtained by letting  $\lambda \rightarrow \infty$  after mass renormalization. I have no proof of the mathematical consistency of this procedure, but the presumption is very strong that it is satisfactory. (It is also strong that a satisfactory form for  $f_+$  can be found.)

## 7 THE PROBLEM OF VACUUM POLARIZATION

In the analysis of the radiative corrections to scattering one type of term was not considered. The potential which we can assume to vary as  $a_\mu \exp(-iq \cdot x)$  creates a pair of electrons (see Fig. 6), momenta  $\mathbf{p}_a, -\mathbf{p}_b$ . This pair then reannihilates, emitting a quantum  $\mathbf{q} = \mathbf{q}_b - \mathbf{q}_a$ , which quantum scatters the original electron from state 1 to state 2. The matrix element for this process (and the others which can be obtained by rearranging the

order in time of the various events) is

$$\begin{aligned} & -(e^2/\pi i)(\tilde{u}_2 \gamma_\mu u_1) \int Sp[(\mathbf{p}_a + \mathbf{q} - m)^{-1} \\ & \times \gamma_\mu (\mathbf{p}_a - m)^{-1} \gamma_\mu] d^4 p_a \mathbf{q}^{-2} C(\mathbf{q}^2) a_\nu. \end{aligned} \quad (30)$$

This is because the potential produces the pair with amplitude proportional to  $a_\nu \gamma_\nu$  the electrons of momenta  $\mathbf{p}_a$ , and  $-(\mathbf{p}_a + \mathbf{q})$  proceed from there to annihilate, producing a quantum (factor  $\gamma_\mu$ ) which propagates (factor  $\mathbf{q}^{-2} C(\mathbf{q}^2)$ ) over to the other electron, by which it is absorbed (matrix element of  $\gamma_\mu$ , between states 1 and 2 of the original electron ( $\tilde{u}_2 \gamma_\mu u_1$ )). All momenta  $\mathbf{p}_a$  and spin states of the virtual electron are admitted, which means the spur and the integral on  $d^4 p_a$  are calculated.

One can imagine that the closed loop path of the positron-electron produces a current

$$4\pi J_\mu a_\nu, \quad (31)$$

which is the source of the quanta which act on the second electron. The quantity

$$\begin{aligned} J_{\mu\nu} = & -(e^2/\pi i) \int Sp[(\mathbf{p} + \mathbf{q} - m)^{-1} \\ & \times \gamma_\mu (\mathbf{p} - m)^{-1} \gamma_\mu] d^4 p, \end{aligned} \quad (32)$$

is then characteristic for this problem of polarization of the vacuum.

One sees at once that  $J_{\mu\nu}$  diverges badly. The modification of  $\delta$  to  $f$  alters the amplitude with which the current  $j_\mu$ , will affect the scattered electron, but it can do nothing to prevent the divergence of the integral (32) and of its effects.

One way to avoid such difficulties is apparent. From one point of view we are considering all routes by which a given electron can get from one region of space-time to another, i.e., from the source of electrons to the apparatus which measures them. From this point of view the closed loop path leading to (32) is unnatural. It might be assumed that the only paths of meaning are those which start from the source and work their way in a continuous path (possibly containing many time reversals) to the detector. Closed loops would be excluded. We have already found that this may be done for electrons moving in a fixed potential.

Such a suggestion must meet several questions, however. The closed loops are a consequence of the usual hole theory in electrodynamics. Among other things, they are required to keep probability conserved. The probability that no pair is produced by a potential is not unity and its deviation from unity arises from the imaginary part of  $J_{\mu\nu}$ . Again, with closed loops

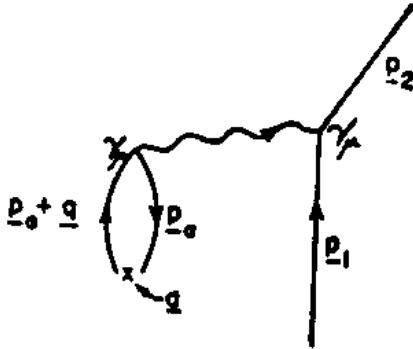


Figure 6: Vacuum polarization effect on scattering, Eq. (30).

excluded, a pair of electrons once created cannot annihilate one another again, the scattering of light by light would be zero, etc. Although we are not experimentally sure of these phenomena, this does seem to indicate that the closed loops are necessary. To be sure, it is always possible that these matters of probability conservation, etc., will work themselves out as simply in the case of interacting particles as for those in a fixed potential. Lacking such a demonstration the presumption is that the difficulties of vacuum polarization are not so easily circumvented.<sup>20</sup>

An alternative procedure discussed in *B* is to assume that the function  $K_+(2, 1)$  used above is incorrect and is to be replaced by a modified function  $K'_+$  having no singularity on the light cone. The effect of this is to provide a convergence factor  $C(\mathbf{p}^2 - m^2)$  for every integral over electron momenta.<sup>21</sup> This will multiply the integrand of (32) by  $C(\mathbf{p}^2 - m^2)C((\mathbf{p} + \mathbf{q})^2 - m^2)$ , since the integral was originally  $\delta(\mathbf{p}_a - \mathbf{p}_b + \mathbf{q})d^4p_a d^4p_b$  and both  $\mathbf{p}_a$  and  $\mathbf{p}_b$  get convergence factors. The integral now converges but the result is unsatisfactory.<sup>22</sup>

One expects the current (31) to be conserved, that is  $q_\mu j_\mu = 0$  or

---

<sup>20</sup>It would be very interesting to calculate the Lamb shift accurately enough to be sure that the 20 megacycles expected from vacuum polarization are actually present.

<sup>21</sup>This technique also makes self-energy and radiationless scattering integrals finite even without the modification of  $\delta_+$  to  $f_+$  for the radiation (and the consequent convergence factor  $C(\mathbf{k}^2)$  for the quanta). See *B*.

<sup>22</sup>Added to the terms given below (33) there is a term  $\frac{1}{4}(\lambda^3 - 2\mu^2 + \frac{1}{3}\mathbf{q}^2)\delta_{\mu\nu}$  for  $C(\mathbf{k}^2) = -\lambda^2(\mathbf{k}^2 - \lambda^2)^{-1}$ , which is not gauge invariant. (In addition the charge renormalization has  $-7/6$  added to the logarithm.)

$q_\mu J_{\mu\nu} = 0$ . Also one expects no current if  $a$  is a gradient, or  $a_\nu = q_\nu$ , times a constant. This leads to the condition  $J_{\mu\nu}q_\nu = 0$  which is equivalent to  $q_\mu J_{\mu\nu} = 0$  since  $J_{\mu\nu}$  is symmetrical. But when the expression (32) is integrated with such convergence factors it does not satisfy this condition. By altering the kernel from  $K$  to another,  $K'$ , which does not satisfy the Dirac equation we have lost the gauge invariance, its consequent current conservation and the general consistency of the theory.

One can see this best by calculating  $J_{\mu\nu}q_\nu$  directly from (32). The expression within the spur becomes  $(\mathbf{p} + \mathbf{q} - m)^{-1}\mathbf{q}(\mathbf{p} - m)^{-1}\gamma_\mu$  which can be written as the difference of two terms:  $(\mathbf{p} - m)^{-1}\gamma_\mu - (\mathbf{p} + \mathbf{q} - m)^{-1}\gamma_\mu$ . Each of these terms would give the same result if the integration  $d^4p$  were without a convergence factor, for the first can be converted into the second by a shift of the origin of  $\mathbf{p}$ , namely  $\mathbf{p}' = \mathbf{p} + \mathbf{q}$ . This does not result in cancelation in (32) however, for the convergence factor is altered by the substitution.

A method of making (32) convergent without spoiling the gauge invariance has been found by Bethe and by Pauli. The convergence factor for light can be looked upon as the result of superposition of the effects of quanta of various masses (some contributing negatively). Likewise if we take the factor  $C(\mathbf{p}^2 - m^2) = -\lambda^2(\mathbf{p}^2 - m^2 - \lambda^2)^{-1}$  so that  $(\mathbf{p}^2 - m^2)^{-1}C(\mathbf{p}^2 - m^2) = (\mathbf{p}^2 - m^2)^{-1} - (\mathbf{p}^2 - m^2 - \lambda^2)^{-1}$  we are taking the difference of the result for electrons of mass  $m$  and mass  $(\lambda^2 + m^2)^{1/2}$ . But we have taken this difference for *each* propagation between interactions with photons. They suggest instead that once created with a certain mass the electron should continue to propagate with this mass through all the potential interactions until it closes its loop. That is if the quantity (32), integrated over some finite range of  $\mathbf{p}$ , is called  $J_{\mu\nu}(m^2)$  and the corresponding quantity over the same range of  $\mathbf{p}$ , but with  $m$  replaced by  $(m^2 + \lambda^2)^{1/2}$  is  $J_{\mu\nu}(m^2 + \lambda^2)$  we should calculate

$$J_{\mu\nu}^P = \int_0^\infty [J_{\mu\nu}(m^2) - J_{\mu\nu}(m^2 + \lambda^2)]G(\lambda)d\lambda, \quad (32')$$

the function  $G(\lambda)$  satisfying  $\int_0^\infty G(\lambda)d\lambda = 1$  and  $\int_0^\infty G(\lambda)\lambda^2d\lambda = 0$ . Then in the expression for  $J_{\mu\nu}^P$  the range of  $\mathbf{p}$  integration can be extended to infinity as the integral now converges. The result of the integration using

this method is the integral on  $d\lambda$  over  $G(\lambda)$  of (see Appendix C)

$$J_{\mu\nu}^P = -\frac{e^2}{\pi}(q_\mu q_\nu - \delta_{\mu\nu}\mathbf{q}^2) \left( -\frac{1}{3} \ln \frac{\lambda^2}{m^2} - \left[ \frac{4m^2 + 2\mathbf{q}^2}{3\mathbf{q}^2} \left( 1 - \frac{\theta}{\tan\theta} \right) - \frac{1}{9} \right] \right), \quad (33)$$

with  $\mathbf{q}^2 = 4m^2 \sin^2 \theta$ .

The gauge invariance is clear, since  $q_\mu(q_\mu q_\nu - \mathbf{q}^2 \delta_{\mu\nu}) = 0$ . Operating (as it always will) on a potential of zero divergence the  $(q_\mu q_\nu - \delta_{\mu\nu}\mathbf{q}^2)a_\nu$ , is simply  $-q^2 a_\mu$ , the D'Alembertian of the potential, that is, the current producing the potential. The term  $-\frac{1}{3}(\ln(\lambda^2/m^2))(q_\mu q_\nu - \mathbf{q}^2 \delta_{\mu\nu})$  therefore gives a current proportional to the current producing the potential. This would have the same effect as a change in charge, so that we would have a difference  $\Delta(e^2)$  between  $e^2$  and the experimentally observed charge,  $e^2 + \Delta(e^2)$ , analogous to the difference between  $m$  and the observed mass. This charge depends logarithmically on the cut-off,  $\Delta(e^2)/e^2 = -(2e^2/3\pi) \ln(\lambda/m)$ . After this renormalization of charge is made, no effects will be sensitive to the cut-off.

After this is done the final term remaining in (33), contains the usual effects<sup>23</sup> of polarization of the vacuum. It is zero for a free light quantum ( $\mathbf{q}^2 = 0$ ). For small  $\mathbf{q}^2$  it behaves as  $(2/15)\mathbf{q}^2$  (adding  $-\frac{1}{5}$  to the logarithm in the Lamb effect). For  $\mathbf{q}^2 > (2m)^2$  it is complex, the imaginary part representing the loss in amplitude required by the fact that the probability that no quanta are produced by a potential able to produce pairs ( $(\mathbf{q}^2)^{1/2} > 2m$ ) decreases with time. (To make the necessary analytic continuation, imagine  $m$  to have a small negative imaginary part, so that  $(1 - \mathbf{q}^2/4m^2 - 1)^{1/2}$  becomes  $-i(\mathbf{q}^2/4m^2 - 1)^{1/2}$  as  $\mathbf{q}^2$  goes from below to above  $4m^2$ . Then  $\theta = \pi/2 + iu$  where  $\sin hu = +(\mathbf{q}^2/4m^2 - 1)^{1/2}$ , and  $-1/\tan\theta = i\tanh u = +i(\mathbf{q}^2 - 4m^2)^{1/2}(\mathbf{q}^2)^{-1/2}$ ).

Closed loops containing a number of quanta or potential interactions larger than two produce no trouble. Any loop with an odd number of interactions gives zero (I, reference 9). Four or more potential interactions give integrals which are convergent even without a convergence factor as is well known. The situation is analogous to that for self-energy. Once the simple problem of a single closed loop is solved there are no further divergence difficulties for more complex processes.<sup>24</sup>

---

<sup>23</sup>E. A. Uehling, Phys. Rev. **48**, 55 (1935), R. Serber, Phys. Rev. **48**, 49 (1935).

<sup>24</sup>There are loops completely without external interactions. For example, a pair is created virtually along with a photon. Next they annihilate, absorbing this photon. Such

## 8 LONGITUDINAL WAVES

In the usual form of quantum electrodynamics the longitudinal and transverse waves are given separate treatment. Alternately the condition  $(\partial A_\mu / \partial x_\mu) \Psi = 0$  is carried along as a supplementary condition. In the present form no such special considerations are necessary for we are dealing with the solutions of the equation  $-\square^2 A_\mu = 4\pi j_\mu$ , with a current  $j_\mu$ , which is conserved  $\partial j_\mu / \partial x_\mu = 0$ . That means at least  $\square^2 (\partial A_\mu / \partial x_\mu) = 0$  and in fact our solution also satisfies  $\partial A_\mu / \partial x_\mu = 0$ .

To show that this is the case we consider the amplitude for emission (real or virtual) of a photon and show that the divergence of this amplitude vanishes. The amplitude for emission for photons polarized in the  $\mu$  direction involves matrix elements of  $\gamma_\mu$ . Therefore what we have to show is that the corresponding matrix elements of  $q_\mu \gamma_\mu = \mathbf{q}$  vanish. For example, for a first order effect we would require the matrix element of  $\mathbf{q}$  between two states  $\mathbf{p}_1$  and  $\mathbf{p}_2 = \mathbf{p}_1 + \mathbf{q}$ . But since  $\mathbf{q} = \mathbf{p}_2 - \mathbf{p}_1$  and  $(\tilde{u}_2 \mathbf{p}_1 u_1) = m(\tilde{u}_2 u_1) = (\tilde{u}_2 \mathbf{p}_2 u_1)$  the matrix element vanishes, which proves the contention in this case. It also vanishes in more complex situations (essentially because of relation (34), below) (for example, try putting  $\mathbf{e}_2 = \mathbf{q}_2$  in the matrix (15) for the Compton Effect).

To prove this in general, suppose  $\mathbf{a}_i, i = 1 \text{ to } N$  are a set of plane wave disturbing potentials carrying momenta  $\mathbf{q}_i$ , (e.g., some may be emissions or absorptions of the same or different quanta) and consider a matrix for the transition from a state of momentum  $\mathbf{p}_0$  to  $\mathbf{p}_N$  such as  $\mathbf{a}_N \Pi_{i=1}^{N-1} (\mathbf{p}_i - m)^{-1} \mathbf{a}_i$ , where  $\mathbf{p}_i = \mathbf{p}_{i-1} + \mathbf{q}_i$  (and in the product, terms with larger  $i$  are written to the left). The most general matrix element is simply a linear combination of these. Next consider the matrix between states  $\mathbf{p}_0$  and  $\mathbf{p}_N + \mathbf{q}$  in a situation in which not only are the  $a_i$ , acting but also another potential  $\mathbf{a} \exp(-iq \cdot x)$  where  $\mathbf{a} = \mathbf{q}$ . This may act previous to all  $\mathbf{a}_i$  in which case it gives  $\mathbf{a}_N \Pi (\mathbf{p}_i + \mathbf{q} - m)^{-1} \mathbf{a}_i (\mathbf{p}_0 + \mathbf{q} - m)^{-1} \mathbf{q}$  which is equivalent to  $+ \mathbf{a}_N \Pi (\mathbf{p}_i + \mathbf{q} - m)^{-1} \mathbf{a}_i$  since  $+ (\mathbf{p}_0 + \mathbf{q} - m)^{-1} \mathbf{q}$  is equivalent to  $(\mathbf{p}_0 + \mathbf{q} - m)^{-1} \times (\mathbf{p}_0 + \mathbf{q} - m)$  as  $\mathbf{p}_0$  is equivalent to  $m$  acting on the initial state. Likewise if it acts after all the potentials it gives  $\mathbf{q} (\mathbf{p}_N - m)^{-1} \mathbf{a}_N \Pi (\mathbf{p}_i - m)^{-1} \mathbf{a}_i$  which is equivalent to  $- \mathbf{a}_N \Pi (\mathbf{p}_i - m)^{-1} \mathbf{a}_i$  since  $\mathbf{p}_N + \mathbf{q} - m$  gives zero on the final state. Or again

---

loops are disregarded on the grounds that they do not interact with anything and are thereby completely unobservable. Any indirect effects they may have via the exclusion principle have already been included.

it may act between the potential  $\mathbf{a}_k$  and  $\mathbf{a}_{k+1}$  for each  $k$ . This gives

$$\sum_{k=1}^{N-1} \mathbf{a}_N \prod_{i=k+1}^{N-1} (\mathbf{p}_i + \mathbf{q} - m)^{-1} \mathbf{a}_i (\mathbf{p}_k + \mathbf{q} - m)^{-1} \\ \times \mathbf{q} (\mathbf{p}_k - m)^{-1} \mathbf{a}_k \prod_{j=1}^{k-1} (\mathbf{p}_j - m)^{-1} \mathbf{a}_j.$$

However,

$$(\mathbf{p}_k + \mathbf{q} - m)^{-1} \mathbf{q} (\mathbf{p}_k - m)^{-1} = (\mathbf{p}_k - m)^{-1} - (\mathbf{p}_k + \mathbf{q} - m)^{-1}, \quad (34)$$

so that the sum breaks into the difference of two sums, the first of which may be converted to the other by the replacement of  $k$  by  $k - 1$ . There remain only the terms from the ends of the range of summation,

$$+ \mathbf{a}_N \prod_{i=1}^{N-1} (\mathbf{p}_i - m)^{-1} \mathbf{a}_i - \mathbf{a}_N \prod_{i=1}^{N-1} (\mathbf{p}_i + \mathbf{q} - m)^{-1} \mathbf{a}_i.$$

These cancel the two terms originally discussed so that the entire effect is zero. Hence any wave emitted will satisfy  $\partial A_\mu / \partial x_\mu = 0$ . Likewise longitudinal waves (that is, waves for which  $A_\mu = \partial \phi / \partial x_\mu$  or  $\mathbf{a} = \mathbf{q}$ ) cannot be absorbed and will have no effect, for the matrix elements for emission and absorption are similar. (We have said little more than that a potential  $A_\mu = \partial \varphi / \partial x_\mu$  has no effect on a Dirac electron since a transformation  $\psi' = \exp(-i\phi)\psi$  removes it. It is also easy to see in coordinate representation using integrations by parts.)

This has a useful practical consequence in that in computing probabilities for transition for unpolarized light one can sum the squared matrix over all four directions rather than just the two special polarization vectors. Thus suppose the matrix element for some process for light polarized in direction  $e_\mu$ , is  $e_\mu M_\mu$ . If the light has wave vector  $q_\mu$ , we know from the argument above that  $q_\mu M_\mu = 0$ . For unpolarized light progressing in the  $z$  direction we would ordinarily calculate  $M_x^2 + M_y^2$ . But we can as well sum  $M_x^2 + M_y^2 + M_z^2 - M_t^2$  for  $q_\mu M_\mu$  implies  $M_t = M_z$  since  $q_t = q_z$  for free quanta. This shows that unpolarized light is a relativistically invariant concept, and permits some simplification in computing cross sections for such light.

Incidentally, the virtual quanta interact through terms like  $\gamma_\mu \dots \gamma_\mu \mathbf{k}^{-2} d^4 k$ . Real processes correspond to poles in the formulae for virtual processes. The pole occurs when  $\mathbf{k}^2 = 0$ , but it looks at first as though in the sum on all four values of  $\mu$ , of  $\gamma_\mu \dots \gamma_\mu$  we would have four kinds of polarization instead of two. Now it is clear that only two perpendicular to  $\mathbf{k}$  are effective.

The usual elimination of longitudinal and scalar virtual photons (leading to an instantaneous Coulomb potential) can of course be performed here too (although it is not particularly useful). A typical term in a virtual transition is  $\gamma_\mu \dots \gamma_\mu \mathbf{k}^{-2} d^4 k$  where the  $\dots$  represent some intervening matrices. Let us choose for the values of  $\mu$ , the time  $t$ , the direction of vector part  $\mathbf{K}$ , of  $\mathbf{k}$ , and two perpendicular directions 1, 2. We shall not change the expression for these two 1, 2 for these are represented by transverse quanta. But we must find  $(\gamma_t \dots \gamma_t) - (\gamma_{\mathbf{K}} \dots \gamma_{\mathbf{K}})$ . Now  $\mathbf{k} = k_4 \gamma_t - K \gamma_{\mathbf{K}}$ , where  $K = (\mathbf{K} \cdot \mathbf{K})^{1/2}$ , and we have shown above that  $\mathbf{k}$  replacing the  $\gamma_\mu$  gives zero.<sup>25</sup> Hence  $K \gamma_{\mathbf{K}}$  is equivalent to  $k_4 \gamma_t$  and

$$(\gamma_t \dots \gamma_t) - (\gamma_{\mathbf{K}} \dots \gamma_{\mathbf{K}}) = ((K^2 - k_4^2)/K^2)(\gamma_t \dots \gamma_t),$$

so that on multiplying by  $\mathbf{k}^{-2} d^4 k = d^4 k (k_4^2 - K^2)^{-1}$  the net effect is  $-(\gamma_t \dots \gamma_t) d^4 k / K^2$ . The  $\gamma_t$  means just scalar waves, that is, potentials produced by charge density. The fact that  $1/K^2$  does not contain  $k_4$  means that  $k_4$  can be integrated first, resulting in an instantaneous interaction, and the  $d^3 \mathbf{K} / K^2$  is just the momentum representation of the Coulomb potential,  $1/r$ .

---

<sup>25</sup>A little more care is required when both  $\gamma_\mu$ 's act on the same particle. Define  $\mathbf{x} = k_4 \gamma_t + K \gamma_{\mathbf{K}}$ , and consider  $(\mathbf{k} \dots \mathbf{x}) + (\mathbf{x} \dots \mathbf{k})$ . Exactly this term would arise if a system, acted on by potential  $\mathbf{x}$  carrying momentum  $-\mathbf{k}$ , is disturbed by an added potential  $\mathbf{k}$  of momentum  $+\mathbf{k}$  (the reversed sign of the momenta in the intermediate factors in the second term  $\mathbf{x} \dots \mathbf{k}$  has no effect since we will later integrate over all  $\mathbf{k}$ ). Hence as shown above the result is zero, but since  $(\mathbf{k} \dots \mathbf{x}) + (\mathbf{x} \dots \mathbf{k}) = k_4^2 (\gamma_t \dots \gamma_t) - K^2 (\gamma_{\mathbf{K}} \dots \gamma_{\mathbf{K}})$  we can still conclude  $(\gamma_{\mathbf{K}} \dots \gamma_{\mathbf{K}}) = k_4^2 K^{-2} (\gamma_t \dots \gamma_t)$ .

## 9 KLEIN GORDON EQUATION

The methods may be readily extended to particles of spin zero satisfying the Klein Gordon equation,<sup>26</sup>

$$\square^2\psi - m^2\psi = i\partial(A_\mu\psi)/\partial x_\mu + iA_\mu\partial\psi/\partial x_\mu - A_\mu A_\mu\psi. \quad (35)$$

The important kernel is now  $I_+(2, 1)$  denned in (I, Eq. (32)). For a free particle, the wave function  $\psi(20$  satisfies  $+\square^2\psi - m^2\psi = 0$ . At a point, 2, inside a space time region it is given by

$$\psi(2) = \int [\psi(1)\partial I_+(2, 1)/\partial x_{1\mu} - (\partial\psi/\partial x_{1\mu})I_+(2, 1)]N_\mu(1)d^3V_1,$$

(as is readily shown by the usual method of demonstrating Green's theorem) the integral being over an entire 3-surface boundary of the region (with normal vector  $N_\mu$ ). Only the positive frequency components of  $\psi$  contribute from the surface preceding the time corresponding to 2, and only negative frequencies from the surface future to 2. These can be interpreted as electrons and positrons in direct analogy to the Dirac case.

The right-hand side of (35) can be considered as a source of new waves and a series of terms written down to represent matrix elements for processes of increasing order. There is only one new point here, the term in  $A_\mu A_\mu$  by which two quanta can act at the same time. As an example, suppose three quanta or potentials,  $a_\mu \exp(-iq_a \cdot x)$ ,  $b_\mu \exp(-iq_b \cdot x)$ , and  $c_\mu \exp(iq_c \cdot x)$  are to act in that order on a particle of original momentum  $p_{0\mu}$ , so that  $\mathbf{p}_a = \mathbf{p}_0 + \mathbf{q}_a$ , and  $\mathbf{p}_b = \mathbf{p}_a + \mathbf{q}_b$ ; the final momentum being  $\mathbf{p}_c = \mathbf{p}_b + \mathbf{q}_c$ .

---

<sup>26</sup>The equations discussed in this section were deduced from the formulation of the Klein Gordon equation given in reference 5, Section 14. The function  $\psi$  in this section has only one component and is not a spinor. An alternative formal method of making the equations valid for spin zero and also for spin 1 is (presumably) by use of the Kemmer-Duffin matrices  $\beta_\mu$  satisfying the commutation relation

$$\beta_\mu\beta_\nu\beta_\sigma + \beta_\sigma\beta_\nu\beta_\mu = \delta_{\mu\nu}\beta_\sigma + \delta_{\sigma\nu}\beta_\mu.$$

If we interpret  $\mathbf{a}$  to mean  $a_\mu\beta_\mu$ , rather than  $a_\mu\gamma_\mu$ , for any  $a_\mu$ , all of the equations in momentum space will remain formally identical to those for the spin 1/2; with the exception of those in which a denominator  $(\mathbf{p} - m)^{-1}$  has been rationalized to  $(\mathbf{p} + m)(\mathbf{p}^2 - m^2)^{-1}$  since  $\mathbf{p}^2$  is no longer equal to a number,  $p \cdot p$ . But  $\mathbf{p}^3$  does equal  $(p \cdot p)\mathbf{p}$  that  $(\mathbf{p} - m)^{-1}$  may now be interpreted as  $(mp + m^2 + \mathbf{p}^2 - p \cdot p)(p \cdot p - m^2)^{-1}$ . This implies that equations in coordinate space will be valid of the function  $K_+(2, 1)$  is given as  $K_+(2, 1) = [(i\nabla_2 + m) - m^{-1}(\nabla_2 + \square_2^2)]iI_+(2, 1)$  with  $\nabla_2 = \beta_\mu\partial/\partial x_{2\mu}$ . This is all in virtue of the fact that the many component wave function  $\psi$  (5 components for spin 0, 10 for spin 1) satisfies  $(i\nabla - m)\psi = \mathbf{a}\psi$  which is formally identical to the Dirac Equation. See W. Pauli, Rev. Mod. Phys. **13**, 203 (1940).

The matrix element is the sum of three terms ( $\mathbf{p}^2 = p_\mu p_\mu$ ) (illustrated in Fig. 7)

$$(p_c \cdot c + p_b \cdot c)(\mathbf{p}_b^2 - m^2)^{-1}(p_b \cdot b + p_a \cdot b) \times (\mathbf{p}_a^2 - m^2)^{-1}(p_a \cdot a + p_0 \cdot a) \\ -(p_c \cdot c + p_b \cdot c)(\mathbf{p}_b^2 - m^2)^{-1}(b \cdot a) - (c \cdot b)(\mathbf{p}_a^2 - m^2)^{-1}(p_a \cdot a + p_0 \cdot a). \quad (36)$$

The first comes when each potential acts through the perturbation  $i\partial(A_\mu\psi)/\partial x_\mu + iA_\mu\partial\psi/\partial x_\mu$ . These gradient operators in momentum space mean respectively the momentum after and before the potential  $A_\mu$  operates. The second term comes from  $b_\mu$  and  $a_\mu$  acting at the same instant and arises from the  $A_\mu A_\mu$  term in (a). Together  $b_\mu$  and  $a_\mu$  carry momentum  $q_{b\mu} + q_{a\mu}$  so that after  $b \cdot a$  operates the momentum is  $\mathbf{p}_0 + \mathbf{q}_a + \mathbf{q}_b$  or  $\mathbf{p}_b$ . The final term comes from  $c_\mu$  and  $b_\mu$  operating together in a similar manner. The term  $A_\mu A_\mu$  thus permits a new type of process in which two quanta can be emitted (or absorbed, or one absorbed, one emitted) at the same time. There is no  $a \cdot c$  term for the order  $a, b, c$  we have assumed. In an actual problem there would be other terms like (36) but with alterations in the order in which the quanta  $a, b, c$  act. In these terms  $a \cdot c$  would appear.

As a further example the self-energy of a particle of momentum  $p_\mu$  is

$$(e^2/2\pi im) \int [(2p - k)_\mu ((\mathbf{p} - \mathbf{k})^2 - m^2)^{-1} \times (2p - k)_\mu - \delta_{\mu\mu}] d^4k \mathbf{k}^{-2} C(\mathbf{k}^2),$$

where the  $\delta_{\mu\mu}$  comes from the  $A_\mu A_\mu$  term and represents the possibility of the simultaneous emission and absorption of the same virtual quantum. This integral without the  $C(\mathbf{k}^2)$  diverges quadratically and would not converge if  $C(\mathbf{k}^2) = -\lambda^2/(k^2 - \lambda^2)$ . Since the interaction occurs through the gradients of the potential, we must use a stronger convergence factor, for example  $C(\mathbf{k}^2) = \lambda^4(k^2 - \lambda^2)^{-2}$ , or in general (17) with  $\int_0^\infty \lambda^2 G(\lambda) d\lambda = 0$ . In this case the self-energy converges but depends quadratically on the cut-off  $\lambda$  and is not necessarily small compared to  $m$ . The radiative corrections to scattering after mass renormalization are insensitive to the cut-off just as for the Dirac equation.

When there are several particles one can obtain Bose statistics by the rule that if two processes lead to the same state but with two electrons exchanged, their amplitudes are to be added (rather than subtracted as for Fermi statistics). In this case equivalence to the second quantization treatment of Pauli and Weisskopf should be demonstrable in a way very much like that given in I (appendix) for Dirac electrons. The Bose statistics mean that the sign of contribution of a closed loop to the vacuum polarization is

the opposite of what it is for the Fermi case (see I). It is ( $\mathbf{p}_b = \mathbf{p}_a + \mathbf{q}$ )

$$J_{\mu\nu} = \frac{e^2}{2\pi im} \int [(p_{b\mu} + p_{a\mu})(p_{b\nu} + p_{a\nu})(\mathbf{p}_a^2 - m^2)^{-1} \\ \times (\mathbf{p}_b^2 - m^2)^{-1} - \delta_{\mu\nu}(\mathbf{p}_a^2 - m^2)^{-1} - \delta_{\mu\nu}(\mathbf{p}_b^2 - m^2)^{-1}] d^4 p_a$$

giving,

$$J_{\mu\nu}^P = \frac{e^2}{\pi} (q_\mu q_\nu - \delta_{\mu\nu} \mathbf{q}^2) \left[ \frac{1}{6} \ln \frac{\lambda^2}{m^2} + \frac{1}{9} - \frac{4m^2 - \mathbf{q}^2}{3\mathbf{q}^2} \left( 1 - \frac{\theta}{\tan\theta} \right) \right],$$

the notation as in (33). The imaginary part for  $(\mathbf{q}^2)^{1/2} > 2m$  is again positive representing the loss in the probability of finding the final state to be a vacuum, associated with the possibilities of pair production. Fermi statistics would give a gain in probability (and also a charge renormalization of opposite sign to that expected).

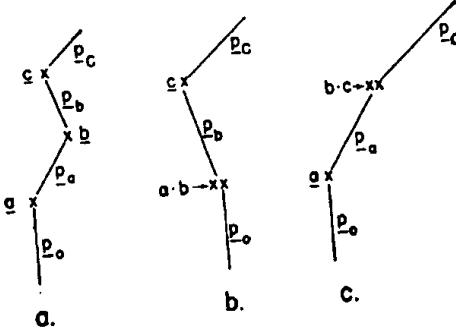


Figure 7: Klein-Gordon particle in three potentials, Eq. (36). The coupling to the electromagnetic field is now, for example,  $p_0 \cdot a + p_a \cdot a$ , and a new possibility arises, (b), of simultaneous interaction with two quanta  $a \cdot b$ . The propagation factor is now  $(p \cdot p - m^2)^{-1}$  for a particle of momentum  $p_\mu$ .

## 10 APPLICATION TO MESON THEORIES

The theories which have been developed to describe mesons and the interaction of nucleons can be easily expressed in the language used here. Calculations, to lowest order in the interactions can be made very easily for the

various theories, but agreement with experimental results is not obtained. Most likely all of our present formulations are quantitatively unsatisfactory. We shall content ourselves therefore with a brief summary of the methods which can be used.

The nucleons are usually assumed to satisfy Dirac's equation so that the factor for propagation of a nucleon of momentum  $\mathbf{p}$  is  $(\mathbf{p} - M)^{-1}$  where  $M$  is the mass of the nucleon (which implies that nucleons can be created in pairs). The nucleon is then assumed to interact with mesons, the various theories differing in the form assumed for this interaction.

First, we consider the case of neutral mesons. The theory closest to electrodynamics is the theory of vector mesons with vector coupling. Here the factor for emission or absorption of a meson is  $g\gamma_\mu$ , when this meson is "polarized" in the  $\mu$  direction. The factor  $g$  the "mesonic charge," replaces the electric charge  $e$ . The amplitude for propagation of a meson of momentum  $\mathbf{q}$  in intermediate states is  $(\mathbf{q}^2 - \mu^2)^{-1}$  (rather than  $\mathbf{q}^{-2}$  as it is for light) where  $\mu$  is the mass of the meson. The necessary integrals are made finite by convergence factors  $C(\mathbf{q}^2 - \mu^2)$  as in electrodynamics. For scalar mesons with scalar coupling the only change is that one replaces the  $\gamma_\mu$  by 1 in emission and absorption. There is no longer a direction of polarization,  $\mu$ , to sum upon. For pseudo-scalar mesons, pseudoscalar coupling replace  $\gamma_\mu$  by  $\gamma_5 = i\gamma_x\gamma_y\gamma_z\gamma_t$ . For example, the self-energy matrix of a nucleon of momentum  $\mathbf{p}$  in this theory is

$$(g^2/\pi i) \int \gamma_5(\mathbf{p} - \mathbf{k} - M)^{-1}\gamma_5 d^4k (\mathbf{k}^2 - \mu^2)^{-1} C(\mathbf{k}^2 - \mu^2).$$

Other types of meson theory result from the replacement of  $\gamma_\mu$ , by other expressions (for example by  $\frac{1}{2}(\gamma_\mu\gamma_\nu - \gamma_\nu\gamma_\mu)$  with a subsequent sum over all  $\mu$  and  $\nu$  for virtual mesons). Scalar mesons with vector coupling result from the replacement of  $\gamma_\mu$  by  $\mu^{-1}\mathbf{q}$  where  $\mathbf{q}$  is the final momentum of the nucleon minus its initial momentum, that is, it is the momentum of the meson if absorbed, or the negative of the momentum of a meson emitted. As is well known, this theory with neutral mesons gives zero for all processes, as is proved by our discussion on longitudinal waves in electrodynamics. Pseudoscalar mesons with pseudo-vector coupling corresponds to  $\gamma_\mu$  being replaced by  $\mu^{-1}\gamma_5\mathbf{q}$  while vector mesons with tensor coupling correspond to using  $(2\mu)^{-1}(\gamma_\mu\mathbf{q} - \mathbf{q}\gamma_\mu)$ . These extra gradients involve the danger of producing higher divergencies for real processes. For example,  $\gamma_5\mathbf{q}$  gives a logarithmically divergent interaction of neutron and electron.<sup>27</sup> Although

---

<sup>27</sup>M. Slotnick and W. Heitler, Phys. Rev. **75**, 1645 (1949).

these divergencies can be held by strong enough convergence factors, the results then are sensitive to the method used for convergence and the size of the cut-off values of  $\lambda$ . For low order processes  $\mu^{-1}\gamma_5\mathbf{q}$  is equivalent to the pseudoscalar interaction  $2M\mu^{-1}\gamma_5$  because if taken between free particle wave functions of the nucleon of momenta  $\mathbf{p}_1$  and  $\mathbf{p}_2 = \mathbf{p}_1 + \mathbf{q}$ , we have

$$(\tilde{u}_2\gamma_5\mathbf{q}u_1) = (\tilde{u}_2\gamma_5(\mathbf{p}_2 - \mathbf{p}_1)u_1) = -(\tilde{u}_2\mathbf{p}_2\gamma_5u_1)$$

$$-(\tilde{u}_2\gamma_5\mathbf{p}_1u_1) = -2M(\tilde{u}_2\gamma_5u_1)$$

since  $\gamma_5$  anticommutes with  $\mathbf{p}_2$  and  $\mathbf{p}_2$  operating on the state 2 equivalent to  $M$  as is 1 on the state 1. This shows that the  $\gamma_5$  interaction is unusually weak in the non-relativistic limit (for example the expected value of  $\gamma_5$  for a free nucleon is zero), but since  $\gamma_5^2 = 1$  is not small, pseudoscalar theory gives a more important interaction in second order than it does in first. Thus the pseudoscalar coupling constant should be chosen to fit nuclear forces including these important second order processes.<sup>28</sup> The equivalence of pseudoscalar and pseudo-vector coupling which holds for low order processes therefore does not hold when the pseudoscalar theory is giving its most important effects. These theories will therefore give quite different results in the majority of practical problems.

In calculating the corrections to scattering of a nucleon by a neutral vector meson field ( $\gamma_\mu$ ) due to the effects of virtual mesons, the situation is just as in electrodynamics, in that the result converges without need for a cut-off and depends only on gradients of the meson potential. With scalar (1) or pseudoscalar ( $\gamma_\mu$ ) neutral mesons the result diverges logarithmically and so must be cut off. The part sensitive to the cut-off, however, is directly proportional to the meson potential. It may thereby be removed by a renormalization of mesonic charge  $g$ . After this renormalization the results depend only on gradients of the meson potential and are essentially independent of cut-off. This is in addition to the mesonic charge renormalization coming from the production of virtual nucleon pairs by a meson, analogous to the vacuum polarization in electrodynamics. But here there is a further difference from electrodynamics for scalar or pseudoscalar mesons in that the polarization also gives a term in the induced current proportional to the meson potential representing therefore an additional renormalization of the *mass of the meson* which usually depends quadratically on the cut-off.

Next consider charged mesons in the absence of an electromagnetic field. One can introduce isotopic spin operators in an obvious way. (Specifically

---

<sup>28</sup>H. A. Bethe, Bull. Am. Phys. Soc. **24**, 3, Z3 (Washington, 1949).

replace the neutral  $\gamma_5$ , say, by  $\tau_i\gamma - 5$  and sum over  $i = 1, 2$ , where  $\tau_1 = \tau_+ + \tau_-$ ,  $\tau_2 = i(\tau_+ - \tau_-)$  and  $\tau_+$  changes neutron to proton ( $\tau_+$  on proton = 0) and  $\tau_-$  changes proton to neutron.) It is just as easy for practical problems simply to keep track of whether the particle is a proton or a neutron on a diagram drawn to help write down the matrix element. This excludes certain processes. For example in the scattering of a negative meson from  $\mathbf{q}_1$  to  $\mathbf{q}_2$  by a neutron, the meson  $\mathbf{q}_2$  must be emitted first (in order of operators, not time) for the neutron cannot absorb the negative meson  $\mathbf{q}_1$  until it becomes a proton. That is, in comparison to the Klein Nishina formula (15), only the analogue of second term (see Fig. 5(b)) would appear in the scattering of negative mesons by neutrons, and only the first term (Fig. 5 (a)) in the neutron scattering of positive mesons.

The source of mesons of a given charge is not conserved, for a neutron capable of emitting negative mesons may (on emitting one, say) become a proton no longer able to do so. The proof that a perturbation  $\mathbf{q}$  gives zero, discussed for longitudinal electromagnetic waves, fails. This has the consequence that vector mesons, if represented by the interaction  $\gamma_\mu$ , would not satisfy the condition that the divergence of the potential is zero. The interaction is to be taken<sup>29</sup> as  $\gamma_\mu - \mu^{-2}q_\mu\mathbf{q}$  in emission and as  $\gamma_\mu$  in absorption if the real emission of mesons with a non-zero divergence of potential is to be avoided. (The correction term  $\mu^{-2}q_\mu\mathbf{q}$  gives zero in the neutral case.) The asymmetry in emission and absorption is only apparent, as this is clearly the same thing as subtracting from the original  $\gamma_\mu \dots \gamma_\mu$ , a term  $\mu^{-2}\mathbf{q} \dots \mathbf{q}$ . That is, if the term  $-\mu^{-2}q_\mu\mathbf{q}$  is omitted the resulting theory

---

<sup>29</sup>The vector meson field potentials  $\varphi_\mu$  satisfy

$$-\partial/\partial x_\nu(\partial\varphi_\mu/\partial x_\nu - \partial\varphi_\nu/\partial x_\mu) - \mu^2\varphi_\mu = -4\pi s_\mu,$$

where  $s_\mu$ , the source for such mesons, is the matrix element of  $\gamma_\mu$  between states of neutron and proton. By taking the divergence  $\partial/\partial x_\mu$  of both sides, conclude that  $\partial\varphi_\nu/\partial x_\nu = 4\pi\mu^{-2}\partial s_\nu/\partial x_\nu$ , so that the original equation can lie rewritten as

$$\square^2\varphi_\mu - \mu^2\varphi_\mu = -4\pi(s_\mu + \mu^{-2}\partial/\partial x_\mu(\partial s_\nu/\partial x_\nu)).$$

The right hand side gives in momentum representation  $\gamma_\mu - \mu^{-2}q_\mu q_\nu \gamma_\nu$  the left yields the  $(\mathbf{q}^2 - \mu^2)^{-1}$  and finally the interaction  $s_\mu\varphi_\mu$  in the Lagrangian gives the  $\gamma_\mu$  on absorption.

Proceeding in this way find generally that particles of spin one can be represented by a four-vector  $u_\mu$  (which, for a free particle of momentum  $q$  satisfies  $q \cdot u = 0$ ). The propagation of virtual particles of momentum  $q$  from state  $\nu$  to  $\mu$  is represented by multiplication by the 4-4 matrix (or tensor)  $P_{\mu\nu} = (\delta_{\mu\nu} - \mu^2 q_\mu q_\nu) \times (q^2 - \mu^2)^{-1}$ . The first-order interaction (from the Proca equation) with an electromagnetic potential  $a \exp(ik \cdot x)$  corresponds to multiplication by the matrix  $E_{\mu\nu} = (q_2 \cdot a + q_1 \cdot a)\delta_{\mu\nu} - q_{2\nu}a_\mu - q_{1\nu}a_\nu$ , where  $q_1$  and  $q_2 = q_1 + k$  are the momenta before and after the interaction. Finally, two potentials  $a, b$  may act simultaneously, with matrix  $E'_{\mu\nu} = -(a \cdot b)\delta_{\mu\nu} + b_\mu a_\nu$

describes a combination of mesons of spin one and spin zero. The spin zero mesons, coupled by vector coupling  $\mathbf{q}$ , are removed by subtracting the term  $\mu^{-2}\mathbf{q} \dots \mathbf{q}$ .

The two extra gradients  $\mathbf{q} \dots \mathbf{q}$  make the problem of diverging integrals still more serious (for example the interaction between two protons corresponding to the exchange of two charged vector mesons depends quadratically on the cut-off if calculated in a straightforward way). One is tempted in this formulation to choose simply  $\gamma_\mu \dots \gamma_\mu$  and accept the admixture of spin zero mesons. But it appears that this leads in the conventional formalism to negative energies for the spin zero component. This shows one of the advantages of the method of second quantization of meson fields over the present formulation. There such errors of sign are obvious while here we seem to be able to write seemingly innocent expressions which can give absurd results. Pseudovector mesons with pseudovector coupling correspond to using  $\gamma_5(\gamma_\mu - \mu^{-2}q_\mu\mathbf{q})$  for absorption and  $\gamma_5\gamma_\mu$  for emission for both charged and neutral mesons.

In the presence of an electromagnetic field, whenever the nucleon is a proton it interacts with the field in the way described for electrons. The meson interacts in the scalar or pseudoscalar case as a particle obeying the Klein-Gordon equation. It is important here to use the method of calculation of Bethe and Pauli, that is, a virtual meson is assumed to have the same "mass" during all its interactions with the electromagnetic field. The result for mass  $\mu$  and for  $(\mu^2 + \lambda^2)^{1/2}$  are subtracted and the difference integrated over the function  $G(\lambda)d\lambda$ . A separate convergence factor is not provided for each meson propagation between electromagnetic interactions, otherwise gauge invariance is not insured. When the coupling involves a gradient, such as  $\gamma - 5\mathbf{q}$  where  $\mathbf{q}$  is the final minus the initial momentum of the nucleon, the vector potential  $\mathbf{A}$  must be subtracted from the momentum of the proton. That is, there is an additional coupling  $\pm\gamma_5\mathbf{A}$  (plus when going from proton to neutron, minus for the reverse) representing the new possibility of a simultaneous emission (or absorption) of meson and photon.

Emission of positive or absorption of negative virtual mesons are represented in the same term, the sign of the charge being determined by temporal relations as for electrons and positrons.

Calculations are very easily carried out in this way to lowest order in  $g^2$  for the various theories for nucleon interaction, scattering of mesons by nucleons, meson production by nuclear collisions and by gamma-rays, nuclear magnetic moments, neutron electron scattering, etc., However, no good agreement with experiment results, when these are available, is obtained. Probably all of the formulations are incorrect. An uncertainty arises since

the calculations are only to first order in  $g^2$ , and are not valid if  $g^2/\hbar c$  is large.

The author is particularly indebted to Professor H. A. Bethe for his explanation of a method of obtaining finite and gauge invariant results for the problem of vacuum polarization. He is also grateful for Professor Bethe's criticisms of the manuscript, and for innumerable discussions during the development of this work. He wishes to thank Professor J. Ashkin for his careful reading of the manuscript.

## APPENDIX

In this appendix a method will be illustrated by which the simpler integrals appearing in problems in electrodynamics can be directly evaluated. The integrals arising in more complex processes lead to rather complicated functions, but the study of the relations of one integral to another and their expression in terms of simpler integrals may be facilitated by the methods given here.

As a typical problem consider the integral (12) appearing in the first order radiationless scattering problem:

$$\int \gamma_\mu (\mathbf{p}_2 - \mathbf{k} - m)^{-1} \mathbf{a}(\mathbf{p}_1 - \mathbf{k} - m)^{-1} \gamma_\mu \mathbf{k}^{-2} d^4 k C(\mathbf{k}^2), \quad (1a)$$

where we shall take  $C(\mathbf{k}^2)$  to be typically  $-\lambda^2(\mathbf{k}^2 - \lambda^2)^{-1}$  and  $d^4 k$  means  $(2\pi)^{-2} dk_1 dk_2 dk_3 dk_4$ . We first rationalize the factors  $(\mathbf{p} - \mathbf{k} - m)^{-1} = (\mathbf{p} - \mathbf{k} + m)((\mathbf{p} - \mathbf{k})^2 - m^2)^{-1}$  obtaining,

$$\begin{aligned} & \int \gamma_\mu (\mathbf{p}_2 - \mathbf{k} + m) \mathbf{a}(\mathbf{p}_1 - \mathbf{k} + m) \gamma_\mu \mathbf{k}^{-2} d^4 k C(\mathbf{k}^2) \\ & \times ((\mathbf{p}_1 - \mathbf{k})^2 - m^2)^{-1} ((\mathbf{p}_2 - \mathbf{k})^2 - m^2)^{-1}. \end{aligned} \quad (2a)$$

The matrix expression may be simplified. It appears to be best to do so after the integrations are performed. Since  $\mathbf{AB} = 2A \cdot B - \mathbf{BA}$  where  $A \cdot B = A_\mu B_\mu$  is a number commuting with all matrices, find, if  $R$  is any expression, and  $\mathbf{A}$  a vector, since  $\gamma_\mu \mathbf{A} = -\mathbf{A} \gamma_\mu + 2A_\mu$ ,

$$\gamma_\mu \mathbf{A} R \gamma_\mu = -\mathbf{A} \gamma_\mu R \gamma_\mu + 2R \mathbf{A}. \quad (3a)$$

Expressions between two  $\gamma_\mu$ 's can be thereby reduced by induction. Particularly useful are

$$\begin{aligned} \gamma_\mu \gamma_\mu &= 4 \\ \gamma_\mu \mathbf{A} \gamma_\mu &= -2\mathbf{A} \\ \gamma_\mu \mathbf{AB} \gamma_\mu &= 2(\mathbf{AB} + \mathbf{BA}) = 4A \cdot B \\ \gamma_\mu \mathbf{ABC} \gamma_\mu &= -2\mathbf{CBA} \end{aligned} \quad (4a)$$

where  $\mathbf{A}, \mathbf{B}, \mathbf{C}$  are any three vector-matrices (i.e., linear combinations of the four  $\gamma_\mu$ s),

In order to calculate the integral in (2a) the integral may be written as the sum of three terms (since  $\mathbf{k} = k_\sigma \gamma_\sigma$ ),

$$\begin{aligned} & \gamma_\mu(\mathbf{p}_2 + m)\mathbf{a}(\mathbf{p}_1 + m)\gamma_\mu J_1 - [\gamma_\mu\gamma_\sigma\mathbf{a}(\mathbf{p}_1 + m)\gamma_\mu \\ & + \gamma_\mu(\mathbf{p}_2 + m)\mathbf{a}\gamma_\sigma\gamma_\mu]J_2 + \gamma_\mu\gamma_\sigma\mathbf{a}\gamma_\tau\gamma_\mu J_3, \end{aligned} \quad (5a)$$

where

$$\begin{aligned} J(1; 2; 3) &= \int (1; k_\sigma; k_\sigma k_\tau) \mathbf{k}^{-2} d^4k C(\mathbf{k}^2) \\ &\times ((\mathbf{p}_2 - \mathbf{k})^2 - m^2)^{-1} ((\mathbf{p}_1 - \mathbf{k})^2 - m^2)^{-1}. \end{aligned} \quad (6a)$$

That is for  $J_1$  the  $(1; k_\sigma; k_\sigma k_\tau)$  is replaced by 1, for  $J_2$  by  $k_\sigma$  and for  $J_3$  by  $k_\sigma k_\tau$ .

More complex processes of the first order involve more factors like  $((\mathbf{p}_3 - \mathbf{k})^2 - m^2)^{-1}$  and a corresponding increase in the number of  $k$ 's which may appear in the numerator, as  $k_\sigma k_\tau k_\nu \dots$ . Higher order processes involving two or more virtual quanta involve similar integrals but with factors possibly involving  $\mathbf{k} + \mathbf{k}'$  instead of just  $\mathbf{k}$ , and the integral extending on  $\mathbf{k}^{-2} d^4k C(\mathbf{k}^2) \mathbf{k}^{-2} d^4k C'(\mathbf{k}'^2)$ . They can be simplified by methods analogous to those used on the first order integrals.

The factors  $(\mathbf{p} - \mathbf{k})^2 - m^2$  may be written

$$(\mathbf{p} - \mathbf{k})^2 - m^2 = \mathbf{k}^2 - 2p \cdot k - \Delta, \quad (7a)$$

where  $\Delta = m^2 - \mathbf{p}^2$ ,  $\Delta_1 = m_1^2 - \mathbf{p}_1^2$ , etc., and we can consider dealing with cases of greater generality in that the different denominators need not have the same value of the mass  $m$ . In our specific problem (6a)  $\mathbf{p}_1^2 = m^2$  that  $\Delta_1 = 0$ , but we desire to work with greater generality.

Now for the factor  $C(\mathbf{k}^2)/\mathbf{k}^2$  we shall use  $-\lambda^2(\mathbf{k}^2 - \lambda^2)^{-1}\mathbf{k}^{-2}$ . This can be written as

$$-\lambda^2/(\mathbf{k}^2 - \lambda^2)\mathbf{k}^2 = \mathbf{k}^{-2}C(\mathbf{k}^2) = -\int_0^{\lambda^2} dL (\mathbf{k}^2 - L)^{-2}. \quad (8a)$$

Thus we can replace  $\mathbf{k}^{-2}C(\mathbf{k}^2)$  by  $(\mathbf{k}^2 - L)^{-2}$  and at the end integrate the result with respect to  $L$  from zero to  $\lambda^2$ . We can for many practical purposes consider  $\lambda^2$  very large relative to  $m^2$  or  $p^2$ . When the original integral converges even without the convergence factor, it will be obvious since the  $L$  integration will then be convergent to infinity. If an infra-red catastrophe exists in the integral one can simply assume quanta have a small mass  $\lambda_{\min}$  and extend the integral on  $L$  from  $\lambda_{\min}^2$  to  $\lambda^2$ , rather than from zero to  $\lambda^2$ .

We then have to do integrals of the form

$$\int (1; k_\sigma; k_\sigma k_\tau d^4k (\mathbf{k}^2 - L)^{-2} (\mathbf{k}^2 - 2p_1 \cdot k - \Delta_1)^{-1} \times (\mathbf{k}^2 - 2p_2 \cdot k - \Delta_2)^{-1}), \quad (9a)$$

where by  $(1; k_\sigma; k_\sigma k_\tau)$  we mean that in the place of this symbol either 1, or  $k_\sigma$  or  $k_\sigma k_\tau$  may stand in different cases. In more complicated problems there may be

more factors  $(\mathbf{k}^2 - 2p_i \cdot k - \Delta_i)^{-1}$  or other powers of these factors (the  $(\mathbf{k}^2 - L)^{-2}$  may be considered as a special case of such a factor with  $\mathbf{p}_i = 0$ ,  $\Delta_i = L$ ) and further factors like  $k_\sigma k_\tau k_\rho \dots$  in the numerator. The poles in all the factors are made definite by the assumption that  $L$ , and the  $\Delta$ 's have infinitesimal negative imaginary parts.

We shall do the integrals of successive complexity by induction. We start with the simplest convergent one, and show

$$\int d^4k (\mathbf{k}^2 - L)^{-3} = (8iL)^{-1}. \quad (10a)$$

For this integral is  $\int (2\pi)^{-2} dk_4 d^3 \mathbf{K} (k_4^2 - \mathbf{K} \cdot \mathbf{K} - L)^{-3}$  where the vector  $\mathbf{K}$ , of magnitude  $K = (\mathbf{K} \cdot \mathbf{K})^{1/2}$  is  $k_1, k_2, k_3$ . The integral on  $k_4$  shows third order poles at  $k_4 = +(K^2 + L)^{1/2}$  and  $k_4 = -(K^2 + L)^{1/2}$ . Imagining, in accordance with our definitions, that  $L$  has a small negative imaginary part only the first is below the real axis. The contour can be closed by an infinite semi-circle below this axis, without change of the value of the integral since the contribution from the semi-circle vanishes in the limit. Thus the contour can be shrunk about the pole  $k_4 = +(K^2 + L)^{1/2}$  and the resulting  $k_4$ , integral is  $-2\pi i$  times the residue at this pole. Writing  $k_4 = (k^2 + L)^{1/2} + \epsilon$  and expanding  $(k_4^2 - K^2 - L)^{-3} = \epsilon^{-3}(\epsilon + 2(K^2 + L)^{1/2})^{-3}$  in powers of  $\epsilon$ , the residue, being the coefficient of the term  $\epsilon^{-1}$ , is seen to be  $6(2(K^2 + L)^{1/2})^{-5}$  so our integral is

$$-(3i/32\pi) \int_0^\infty 4\pi K^2 dK (K^2 + L)^{-5/2} = (3/8i)(1/3L)$$

establishing (10a).

We also have  $\int k_\sigma d^4k (\mathbf{k}^2 - L)^{-3} = 0$  from the symmetry in the  $k$  space. We write these results as

$$(8i) \int (1; k_\sigma) d^4k (\mathbf{k}^2 - L)^{-3} = (1; 0)L^{-1}, \quad (11a)$$

where in the brackets  $(1; k_\sigma)$  and  $(1; 0)$  corresponding entries are to be used.

Substituting  $\mathbf{k} = \mathbf{k}' - \mathbf{p}$  in (11a) and calling  $L - p^2 = \Delta$  shows that

$$(8i) \int (1; k_\sigma) d^4k (\mathbf{k}^2 - 2p \cdot k - \Delta)^{-3} = (1; p_\sigma)(p^2 + \Delta)^{-1}. \quad (12a)$$

By differentiating both sides of (12a) with respect to  $\Delta$  or with respect to  $p_\tau$  there follows directly

$$\begin{aligned} & (24i) \int (1; k_\sigma; k_\sigma k_\tau) d^4k (\mathbf{k}^2 - 2p \cdot k - \Delta)^{-4} \\ &= -(1; p_\sigma; p_\sigma p_\tau - \frac{1}{2}\delta_{\sigma\tau}(p^2 + \Delta))(p^2 + \Delta)^{-2}. \end{aligned} \quad (13a)$$

Further differentiations give directly successive integrals including more  $k$  factors in the numerator and higher powers of  $(\mathbf{k}^2 - 2p \cdot k - \Delta)$  in the denominator.

The integrals so far only contain one factor in the denominator. To obtain results for two factors we make use of the identity

$$a^{-1}b^{-1} = \int_0^1 dx(ax + b(1-x))^{-2}, \quad (14a)$$

(suggested by some work of Schwinger's involving Gaussian integrals). This represents the product of two reciprocals as a parametric integral over one and will therefore permit integrals with two factors to be expressed in terms of one. For other powers of  $a, b$  we make use of all of the identities, such as

$$a^{-2}b^{-1} = \int_0^1 2xdx(ax + b(1-x))^{-3}, \quad (15a)$$

deducible from (14a) by successive differentiations with respect to  $a$  or  $b$ . To perform an integral, such as

$$(8i) \int (1; k_\sigma) d^4k (\mathbf{k}^2 - 2p_1 \cdot k - \Delta_1)^{-2} (\mathbf{k}^2 - 2p_2 \cdot k - \Delta_2)^{-1}, \quad (16a)$$

write, using (15a),

$$(\mathbf{k}^2 - 2p_1 \cdot k - \Delta_1)^{-2} (\mathbf{k}^2 - 2p_2 \cdot k - \Delta_2)^{-1} = \int_0^1 2xdx (\mathbf{k}^2 - 2p_x \cdot k - \Delta_x)^{-3},$$

where

$$\mathbf{p}_x = x\mathbf{p}_1 + (1-x)\mathbf{p}_2 \quad \text{and} \quad \Delta_x = x\Delta_1 + (1-x)\Delta_2, \quad (17a)$$

(note that  $\Delta_x$  is *not* equal to  $m^2 - \mathbf{p}_x^2$ ) so that the expression (16a) is  $(8i) \int_0^1 2xdx (1; k_\sigma) d^4k (\mathbf{k}^2 - 2p_x \cdot k - \Delta_x)^{-3}$  which may now be evaluated by (12a) and is

$$(16a) = \int_0^1 (1; p_{x\sigma}) 2xdx (\mathbf{p}_x^2 + \Delta_x)^{-1}, \quad (18a)$$

where,  $\mathbf{p}_x, \Delta_x$  are given in (17a). The integral in (18a) is elementary, being the integral of ratio of polynomials, the denominator of second degree in  $x$ . The general expression although readily obtained is a rather complicated combination of roots and logarithms.

Other integrals can be obtained again by parametric differentiation. For example differentiation of (16a), (18a) with respect to  $\Delta_2$  or  $p_{2\tau}$  gives

$$\begin{aligned} & (8i) \int (1; k_\sigma; k_\sigma k_\tau) d^4k (\mathbf{k}^2 - 2p_1 \cdot k - \Delta_1)^{-2} (\mathbf{k}^2 - 2p_2 \cdot k - \Delta_2)^{-2} \\ &= - \int_0^1 (1; p_{x\sigma}; p_{x\sigma} p_{x\tau} - \frac{1}{2} \delta_{\sigma\tau} (x^2 \mathbf{p}^2 + \Delta_x)) \times 2x(1-x) dx (\mathbf{p}_x^2 + \Delta_x)^{-2}, \end{aligned} \quad (19a)$$

again leading to elementary integrals.

As an example, consider the case that the second factor is just  $(k^2 - L)^{-2}$  and in the first put  $\mathbf{p}_1 = \mathbf{p}$ ,  $\Delta_1 = \Delta$ . Then  $\mathbf{p}_x = x\mathbf{p}$ ,  $\Delta_x = x\Delta + (1-x)L$ . There results

$$(8i) \int (1; k_\sigma; k_\sigma k_\tau) d^4 k (\mathbf{k}^2 - L)^{-2} (\mathbf{k}^2 - 2p \cdot k - \Delta)^{-2} \\ = - \int_0^1 (1; xp_\sigma; x^2 p_\sigma p_\tau - \frac{1}{2} \delta_{\sigma\tau} (x^2 \mathbf{p}^2 + \Delta_x)) \times 2x(1-x) dx (x^2 \mathbf{p}^2 + \Delta_x)^{-2}. \quad (20a)$$

Integrals with three factors can be reduced to those involving two by using (14a) again. They, therefore, lead to integrals with two parameters (e.g., see application to radiative correction to scattering below).

The methods of calculation given in this paper are deceptively simple when applied to the lower order processes. For processes of increasingly higher orders the complexity and difficulty increases rapidly, and these methods soon become impractical in their present form.

## A. Self-Energy

The self-energy integral (19) is

$$(e^2/\pi i) \int \gamma_\mu (\mathbf{p} - \mathbf{k} - m)^{-1} \gamma_\mu \mathbf{k}^{-2} d^4 k C(\mathbf{k}^2), \quad (19)$$

so that it requires that we find (using the principle of (8a)) the integral on  $L$  from 0 to  $\lambda^2$  of

$$\int \gamma_\mu (\mathbf{p} - \mathbf{k} + m) \gamma_\mu d^4 k (\mathbf{k}^2 - L)^{-2} (\mathbf{k}^2 - 2p \cdot k)^{-1},$$

since  $(\mathbf{p} - \mathbf{k})^2 - m^2 = \mathbf{k}^2 - 2p \cdot k$ , as  $\mathbf{p}^2 = m^2$ . This is of the form (16a) with  $\Delta_1 = L$ ,  $\mathbf{p}_1 = 0$ ,  $\Delta_2 = 0$ ,  $\mathbf{p}_2 = \mathbf{p}$  so that (18a) gives, since  $\mathbf{p}_x = (1-x)\mathbf{p}$ ,  $\Delta_x = xL$ ,

$$(8i) \int (1; k_\sigma) d^4 k (\mathbf{k}^2 - L)^{-2} (\mathbf{k}^2 - 2p \cdot k)^{-1} = \int_0^1 (1; (1-x)p_\sigma) 2x dx ((1-x)^2 m_x^2 L)^{-1},$$

or performing the integral on  $L$ , as in (8),

$$(8i) \int (1; k_\sigma) d^4 k \mathbf{k}^{-2} C(\mathbf{k}^2) (\mathbf{k}^2 - 2p \cdot k)^{-1} = \int_0^1 (1; (1-x)p_\sigma) 2dx \ln \frac{x\lambda^2 + (1-x)^2 m^2}{(1-x)^2 m^2}.$$

Assuming now that  $\lambda^2 \gg m^2$  we neglect  $(1-x)^2 m^2$  relative to  $x\lambda^2$  in the argument of the logarithm, which then becomes  $(\lambda^2/m^2)(x/(1-x)^2)$ . Then since  $\int_0^1 dx \ln(x(1-x)^{-2}) = 1$  and  $\int_0^1 (1-x) dx \ln(x(1-x)^{-2}) = -(1/4)$  find

$$(8i) \int (1; k_\sigma) \mathbf{k}^{-2} C(\mathbf{k}^2) d^4 k (\mathbf{k}^2 - 2p \cdot k)^{-1} = \left( 2 \ln \frac{\lambda^2}{m^2} + 2; p_\sigma \left( \ln \frac{\lambda^2}{m^2} - \frac{1}{2} \right) \right),$$

so that substitution into (19) (after the  $(\mathbf{p} - \mathbf{k} - m)^{-1}$  in (19) is replaced by  $(\mathbf{p} - \mathbf{k} + m)(\mathbf{k}^2 - 2\mathbf{p} \cdot \mathbf{k})^{-1}$ ) gives

$$\begin{aligned}(19) &= (e^2/8\pi)\gamma_\mu[(\mathbf{p} + m)(2\ln(\lambda^2/m^2) + 2) - \mathbf{p}(\ln(\lambda^2/m^2) - \frac{1}{2})]\gamma_\mu \\ &= (e^2/8\pi)[8m(\ln(\lambda^2/m^2) + 1) - \mathbf{p}(2\ln(\lambda^2/m^2) + 5)],\end{aligned}\quad (20)$$

using (4a) to remove the  $\gamma_\mu$ 's. This agrees with Eq. (20) of the text, and gives the self-energy (21) when  $\mathbf{p}$  is replaced by  $m$ .

## B. Corrections to Scattering

The term (12) in the radiationless scattering, after rationalizing the matrix denominators and using  $p_1^2 = p_2^2 = m^2$  requires the integrals (9a), as we have discussed. This is an integral with three denominators which we do in two stages. First the factors  $(\mathbf{k}^2 - 2p_1 \cdot k)$  and  $(\mathbf{k}^2 - 2p_2 \cdot k)$  are combined by a parameter  $y$ ;

$$(\mathbf{k}^2 - 2p_1 \cdot k)^{-1}(\mathbf{k}^2 - 2p_2 \cdot k)^{-1} = \int_0^1 dy (\mathbf{k}^2 - 2p_y \cdot k)^{-2},$$

from (14a) where

$$p_y = yp_1 + (1-y)p_2. \quad (21a)$$

We therefore need the integrals

$$(8i) \int (1; k_\sigma; k_\sigma k_\tau) d^4k (\mathbf{k}^2 - L)^{-2} (\mathbf{k}^2 - 2p_y \cdot k)^{-2}, \quad (22a)$$

which we will then integrate with respect to  $y$  from 0 to 1. Next we do the integrals (22a) immediately from (20a) with  $p = p_y$ ,  $\Delta = 0$ :

$$\begin{aligned}(22a) &= - \int_0^1 \int_0^1 (1; xp_{y\sigma}; x^2 p_{y\sigma} p_{y\tau} \\ &\quad - \frac{1}{2} \delta_{\sigma\tau} (x^2 p_y^2 + (1-x)L)) 2x(1-x) dx (x^2 p_y^2 + L(1-x))^{-2} dy.\end{aligned}$$

We now turn to the integrals on  $L$  as required in (8a). The first term, (1), in  $(1; k_\sigma; k_\sigma k_\tau)$  gives no trouble for large  $L$ , but if  $L$  is put equal to zero there results  $x^{-2} p_y^{-2}$  which leads to a diverging integral on  $x$  as  $x \rightarrow 0$ . This infra-red catastrophe is analyzed by using  $\lambda_{\min}^2$  for the lower limit of the  $L$  integral. For the last term the upper limit of  $L$  must be kept as  $\lambda^2$ . Assuming  $\lambda_{\min}^2 \ll p_y^2 \ll \lambda^2$  the  $x$  integrals which remain are trivial, as in the self-energy case. One finds

$$\begin{aligned}-(8i) \int &(\mathbf{k}^2 - \lambda_{\min}^2)^{-1} d^4k C(\mathbf{k}^2 - \lambda_{\min}^2)(\mathbf{k}^2 - 2p_1 \cdot k)^{-1} (\mathbf{k}^2 - 2p_2 \cdot k)^{-1} \\ &= \int_0^1 p_y^{-2} dy \ln(p_y^2/\lambda_{\min}^2)\end{aligned}\quad (23a)$$

$$\begin{aligned}
& -(8i) \int k_\sigma \mathbf{k}^{-2} d^4 k C(\mathbf{k}^2) (\mathbf{k}^2 - 2\mathbf{p}_1 \cdot \mathbf{k})^{-1} (\mathbf{k}^2 - 2\mathbf{p}_2 \cdot \mathbf{k})^{-1} \\
& = 2 \int_0^1 p_{y\sigma} p_y^{-2} dy,
\end{aligned} \tag{24a}$$

$$\begin{aligned}
& -(8i) \int k_\sigma k_\tau \mathbf{k}^{-2} d^4 k C(\mathbf{k}^2) (\mathbf{k}^2 - 2\mathbf{p}_1 \cdot \mathbf{k})^{-1} (\mathbf{k}^2 - 2\mathbf{p}_2 \cdot \mathbf{k})^{-1} \\
& = \int_0^1 p_{y\sigma} p_{y\tau} p_y^{-2} dy - \frac{1}{2} \delta_{\sigma\tau} \int_0^1 dy \ln(\lambda^2 p_y^{-2}) + \frac{1}{4} \delta_{\sigma\tau}.
\end{aligned} \tag{25a}$$

The integrals on  $y$  give,

$$\int_0^1 p_y^{-2} dy \ln(p_y^2 \lambda_{\min}^{-2}) = 4(m^2 \sin 2\theta)^{-1} \left[ \theta \ln(m \lambda_{\min}^{-1}) - \int_0^\theta \alpha \tan \alpha d\alpha \right], \tag{26a}$$

$$\int_0^1 p_{y\sigma} p_y^{-2} dy = \theta(m^2 \sin 2\theta)^{-1} (p_{1\sigma} + P_{2\sigma}), \tag{27a}$$

$$\begin{aligned}
& \int_0^1 p_{y\sigma} p_{y\tau} p_y^{-2} dy = \theta(2m^2 \sin 2\theta)^{-1} (p_{1\sigma} + p_{1\tau} (p_{2\sigma} + p_{2\tau})) \\
& + \mathbf{q}^{-2} q_\sigma q_\tau (1 - \theta \operatorname{ctn} \theta),
\end{aligned} \tag{28a}$$

$$\int_0^1 dy \ln(\lambda^2 p_y^{-2}) = \ln(\lambda^2/m^2) + 2(1 - \theta \operatorname{ctn} \theta). \tag{29a}$$

These integrals on  $y$  were performed as follows. Since  $\mathbf{p}_2 = \mathbf{p}_1 + \mathbf{q}$  where  $\mathbf{q}$  is the momentum carried by the potential, it follows from  $\mathbf{p}_2^2 = \mathbf{p}_1^2 = m^2$  that  $2\mathbf{p}_1 \cdot \mathbf{q} = -\mathbf{q}^2$  so that since  $\mathbf{p}_y = \mathbf{p}_1 + \mathbf{q}(1-y)$ ,  $\mathbf{p}_y^2 = m^2 - \mathbf{q}^2 y(1-y)$ . The substitution  $2y-1 = \tan \theta$  where  $\theta$  is defined by  $4m^2 \sin^2 \theta = \mathbf{q}^2$  is useful for it means  $\mathbf{p}_y^2 = m^2 \sec^2 \alpha / \sec^2 \theta$  and  $\mathbf{p}_y^{-2} dy = (m^2 \sin 2\theta)^{-1} d\alpha$  where  $\alpha$  goes from  $-\theta$  to  $+\theta$ .

These results are substituted into the original scattering formula (2a), giving (22). It has been simplified by frequent use of the fact that  $\mathbf{p}_1$  operating on the initial state is  $m$  and likewise  $\mathbf{p}_2$  when it appears at the left is replacable by  $m$ . (Thus, to simplify:

$$\begin{aligned}
& \gamma_\mu \mathbf{p}_2 \mathbf{a} \mathbf{p}_1 \gamma_\mu = -2 \mathbf{p}_1 \mathbf{a} \mathbf{p}_2 \quad \text{by (4a),} \\
& = -2(\mathbf{p}_2 - \mathbf{q}) \mathbf{a} (\mathbf{p}_1 + \mathbf{q}) = -2(m - \mathbf{q}) \mathbf{a} (m + \mathbf{q}).
\end{aligned}$$

A term like  $\mathbf{q} \mathbf{a} \mathbf{q} = -q^2 \mathbf{a} + 2(a \cdot q) \mathbf{q}$  is equivalent to just  $-q^2 \mathbf{a}$  since  $\mathbf{q} = \mathbf{p}_2 - \mathbf{p}_1 = m - m$  has zero matrix element.) The renormalization term requires the corresponding integrals for the special case  $\mathbf{q} = 0$ .

## C. Vacuum Polarization

The expressions (32) and (32') for  $J_{\mu\nu}$  in the vacuum polarization problem require the calculation of the integral

$$\begin{aligned} J_{\mu\nu}(m^2) &= \frac{e^2}{\pi i} \int Sp[\gamma_\mu(\mathbf{p} - \frac{1}{2}\mathbf{q} + m)\gamma_\nu\mathbf{p} + \frac{1}{2}\mathbf{q} + m]d^4p \\ &\times ((\mathbf{p} - \frac{1}{2}\mathbf{q})^2 - m^2)^{-1}((\mathbf{p} + \frac{1}{2}\mathbf{q})^2 - m^2)^{-1}, \end{aligned} \quad (32)$$

where we have replaced  $\mathbf{p}$  by  $\mathbf{p} - \frac{1}{2}\mathbf{q}$  to simplify the calculation somewhat. We shall indicate the method of calculation by studying the integral,

$$I(m^2) = \int p_\sigma p_\tau d^4p ((\mathbf{p} - \frac{1}{2}\mathbf{q})^2 - m^2)^{-1}((\mathbf{p} + \frac{1}{2}\mathbf{q})^2 - m^2)^{-1}.$$

The factors in the denominator,  $\mathbf{p}^2 - p \cdot q - m^2 + \frac{1}{4}\mathbf{q}^2$  and  $\mathbf{p}^2 + p \cdot q - m^2 + \frac{1}{4}\mathbf{q}^2$  are combined as usual by (8a) but for symmetry we substitute  $x = \frac{1}{2}(1 + \eta)$ ,  $(1 - x) = \frac{1}{2}(1 - \eta)$  and integrate  $\eta$  from  $-1$  to  $+1$ :

$$I(m^2) = \int_{-1}^{+1} p_\sigma p_\tau d^4p (\mathbf{p}^2 - \eta p \cdot q - m^2 + \frac{1}{4}\mathbf{q}^2)^{-2} d\eta/2. \quad (30a)$$

But the integral on  $\mathbf{p}$  will not be found in our list for it is badly divergent. However, as discussed in Section 7, Eq. (34') we do not wish  $I(m^2)$  but rather  $\int_0^\infty [I(m^2) - I(m^2 + \lambda^2)]G(\lambda)d\lambda$ . We can calculate the difference  $I(m^2) - I(m^2 + \lambda^2)$  by first calculating the derivative  $I'(m^2 + L)$  of  $I$  with respect to  $m^2$  at  $m^2 + L$  and later integrating  $L$  from zero to  $\lambda^2$ . By differentiating (30a), with respect to  $m^2$  find,

$$I'(m^2 + L) = \int_{-1}^{+1} p_\sigma p_\tau d^4p (\mathbf{p}^2 - \eta p \cdot q - m^2 - L + \frac{1}{4}\mathbf{q}^2)^{-3} d\eta.$$

This still diverges, but we can differentiate again to get

$$\begin{aligned} I''(m^2 + L) &= 3 \int_{-1}^{+1} p_\sigma p_\tau d^4p (\mathbf{p}^2 - \eta p \cdot q - m^2 - L + \frac{1}{4}\mathbf{q}^2)^{-4} d\eta \\ &= -(8i)^{-1} \int_{-1}^{+1} (\frac{1}{4}\eta^2 q_\sigma q_\tau D^{-2} - \frac{1}{2}\delta_{\sigma\tau} D^{-1}) d\eta \end{aligned} \quad (31a)$$

(where  $D = \frac{1}{4}(\eta^2 - 1)\mathbf{q}^2 + m^2 + L$ ), which now converges and has been evaluated by (13a) with  $\mathbf{p} = \frac{1}{2}\eta\mathbf{q}$  and  $\Delta = m^2 + L - \frac{1}{4}\mathbf{q}^2$ . Now to get  $I'$  we may integrate

$I''$  with respect to  $L$  as an indefinite integral and we may choose any convenient arbitrary constant. This is because a constant  $C$  in  $I'$  will mean a term  $-C\lambda^2$  in  $I(m^2) - I(m^2 + \lambda^2)$  which vanishes since we will integrate the results times  $G(\lambda)d\lambda$  and  $\int_0^\infty \lambda^2 G(\lambda)d\lambda = 0$ . This means that the logarithm appearing on integrating  $L$  in (31a) presents no problem. We may take

$$I'(m^2 + L) = (8i)^{-1} \int_{-1}^{+1} [\frac{1}{4}\eta^2 q_\sigma q_\tau D^{-1} + \frac{1}{2}\delta_{\sigma\tau} \ln D] d\eta + C\delta_{\sigma\tau},$$

a subsequent integral on  $L$  and finally on  $\eta$  presents no new problems. There results

$$\begin{aligned} & -(8i) \int p_\sigma p_\tau d^4 p ((\mathbf{p} - \frac{1}{2}\mathbf{q})^2 - m^2)^{-1} ((\mathbf{p} + \frac{1}{2}\mathbf{q})^2 - m^2)^{-1} \\ &= (q_\sigma q_\tau - \delta_{\sigma\tau} \mathbf{q}^2) \left[ \frac{1}{9} - \frac{4m^2 - \mathbf{q}^2}{3\mathbf{q}^2} \left( 1 - \frac{\theta}{\tan\theta} \right) + \frac{1}{6} \ln \frac{\lambda^2}{m^2} \right] \\ & \quad + \delta_{\sigma\tau} [(\lambda^2 + m^2) \ln(\lambda^2 m^{-2} + 1) - C' \lambda^2] \end{aligned} \quad (32a)$$

where we assume  $\lambda^2 \gg m^2$  and have put some terms into the arbitrary constant  $C'$  which is independent of  $\lambda^2$  (but in principle could depend on  $\mathbf{q}^2$  and which drops out in the integral on  $G(\lambda)d\lambda$ ). We have set  $\mathbf{q}^2 = 4m^2 \sin^2 \theta$ .

In a very similar way the integral with  $m^2$  in the numerator can be worked out. It is, of course, necessary to differentiate this  $m^2$  also when calculating  $I'$  and  $I''$ . There results

$$\begin{aligned} & -(8i) \int m^2 d^4 p ((\mathbf{p} - \frac{1}{2}\mathbf{q})^2 - m^2)^{-1} ((\mathbf{p} + \frac{1}{2}\mathbf{q})^2 - m^2)^{-1} \\ &= 4m^2 (1 - \theta \operatorname{ctn}\theta) - \mathbf{q}^2/3 + 2(\lambda^2 + m^2) \ln(\lambda^2 m^{-2} + 1) - C'' \lambda^2, \end{aligned} \quad (33a)$$

with another unimportant constant  $C''$ . The complete problem requires the further integral,

$$\begin{aligned} & -(8i) \int (1; p_\sigma) d^4 p ((\mathbf{p} - \frac{1}{2}\mathbf{q})^2 - m^2)^{-1} ((\mathbf{p} + \frac{1}{2}\mathbf{q})^2 - m^2)^{-1} \\ &= (1, 0) (4(1 - \theta \operatorname{ctn}\theta) + 2 \ln(\lambda^2 m^{-2})). \end{aligned} \quad (34a)$$

The value of the integral (34a) times  $m^2$  differs from (33a), of course, because the results on the right are not actually the integrals on the left, but rather equal their actual value minus their value for  $m^2 = m^2 + \lambda^2$ .

Combining these quantities, as required by (32), dropping the constants  $C', C''$  and evaluating the spur gives (33). The spurs are evaluated in the usual way, noting that the spur of any odd number of  $\gamma$  matrices vanishes and  $S_p(AB) = S_p(BA)$  for arbitrary  $A, B$ . The  $S_p(1) = 4$  and we also have

$$\frac{1}{2} S_p[(p_1 + m_1)(\mathbf{p}_2 - m_2)] = p_1 \cdot p_2 - m_1 m_2, \quad (35a)$$

$$\frac{1}{2} S_p [(\mathbf{p}_1 + m_1)(\mathbf{p}_2 - m_2)(\mathbf{p}_4 - m_4)] = (p_1 \cdot p_2 - m_1 m_2)(p_3 \cdot p_4 - m_3 m_4) \\ - (p_1 \cdot p_3 - m_1 m_3)(p_2 \cdot p_4 - m_2 m_4) + (P - 1 \cdot p_4 - m_1 m_4)(p_2 \cdot p_3 - m_2 m_3), \quad (36a)$$

where  $\mathbf{p}_i, m_i$  are arbitrary four-vectors and constants.

It is interesting that the terms of order  $\lambda^2 \ln \lambda^2$  go out, so that the charge renormalization depends only logarithmically on  $\lambda^2$ . This is not true for some of the meson theories. Electrodynamics is suspiciously unique in the mildness of its divergence.

## D. More Complex Problems

Matrix elements for complex problems can be set up in a manner analogous to that used for the simpler cases. We give three illustrations; higher order corrections to the Møller scattering, to the Compton scattering, and the interaction of a neutron

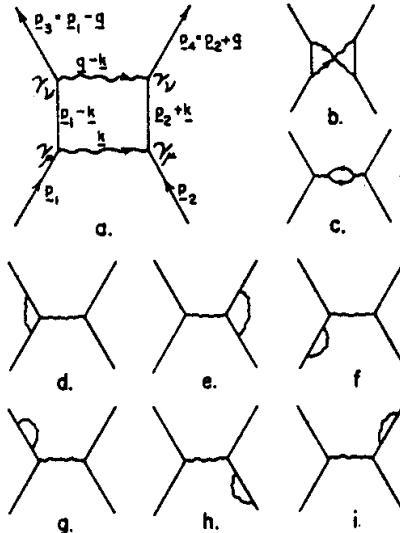


Figure 8: The interaction between two electrons to order  $(e^2/\hbar c)^2$ . One adds the contribution of every figure involving two virtual quanta, Appendix D.

with an electromagnetic field.

For the Møller scattering, consider two electrons, one in state  $u_1$  of momentum  $\mathbf{p}_1$  and the other in state  $u_2$  of momentum  $\mathbf{p}_2$ . Later they are found in states  $u_3, \mathbf{p}_3$  and  $u_4, \mathbf{p}_4$ . This may happen (first order in  $e^2/\hbar c$ ) because they exchange a quantum of momentum  $\mathbf{q} = \mathbf{p}_1 - \mathbf{p}_3 = \mathbf{p}_4 - \mathbf{p}_2$  in the manner of Eq. (4) and

Fig. 1. The matrix element for this process is proportional to (translating (4) to momentum space)

$$(\tilde{u}_4\gamma_\mu u_2)(\tilde{u}_3\gamma_\mu u_1)\mathbf{q}^{-2}. \quad (37a)$$

We shall discuss corrections to (37a) to the next order in  $e^2/\hbar c$ . (There is also the possibility that it is the electron at 2 which finally arrives at 3, the electron at 1 going to 4 through the exchange of quantum of momentum  $\mathbf{p}_3 - \mathbf{p}_2$ . The amplitude for this process,  $(\tilde{u}_4\gamma_\mu u_1)(\tilde{u}_3\gamma_\mu u_2)(\mathbf{p}_3 - \mathbf{p}_2)^{-2}$  must be subtracted from (37a) in accordance with the exclusion principle. A similar situation exists to each order so that we need consider in detail only the corrections to (37a), reserving to the last the subtraction of the same terms with 3, 4 exchanged.) One reason that (37a) is modified is that two quanta may be exchanged, in the manner of Fig. 8a. The total matrix element for all exchanges of this type is

$$(e^2/\pi i) \int (\tilde{u}_3\gamma_\nu(\mathbf{p}_1 - \mathbf{k} - m)^{-1}\gamma_\mu u_1)(\tilde{u}_4\gamma_\nu(\mathbf{p}_2 + \mathbf{k} - m)^{-1}\gamma_\mu u_2) \cdot \mathbf{k}^{-2}(\mathbf{q} - \mathbf{k})^{-2}d^4k, \quad (38a)$$

as is clear from the figure and the general rule that electrons of momentum  $\mathbf{p}$  contribute in amplitude  $(\mathbf{p} - m)^{-1}$  between interactions  $\gamma_\mu$  and that quanta of momentum  $\mathbf{k}$  contribute  $\mathbf{k}^{-2}$ . In integrating on  $d^4k$  and summing over  $\mu$  and  $\nu$ , we add all alternatives of the type of Fig. 8a. If the time of absorption,  $\gamma_\mu$ , of the quantum  $\mathbf{k}$  by electron 2 is later than the absorption,  $\gamma_\mu$ , of  $\mathbf{q} - \mathbf{k}$ , this corresponds to the virtual state  $\mathbf{p}_2 + \mathbf{k}$  being a positron (so that (38a) contains over thirty terms of the conventional method of analysis).

In integrating over all these alternatives we have considered all possible distortions of Fig. 8a which preserve the order of events along the trajectories. We have not included the possibilities corresponding to Fig. 8b, however. Their contribution is

$$(e^2/\pi i) \int (\tilde{u}_3\gamma_\nu(\mathbf{p}_1 - \mathbf{k} - m)^{-1}\gamma_\nu u_1) \times (\tilde{u}_4\gamma_\mu(\mathbf{p}_2 + \mathbf{q} - \mathbf{k} - m)^{-1}\gamma_\nu u_2)\mathbf{k}^{-2}(\mathbf{q} - \mathbf{k})^{-2}d^4k, \quad (39a)$$

as is readily verified by labeling the diagram. The contributions of all possible ways that an event can occur are to be added. This means that one adds with equal weight the integrals corresponding to each topologically distinct figure.

To this same order there are also the possibilities of Fig. 8d which give

$$(e^2/\pi i) \int (\tilde{u}_3\gamma_\nu(\mathbf{p}_2 - \mathbf{k} - m)^{-1}\gamma_\mu(\mathbf{p}_1 - \mathbf{k} - m)^{-1}\gamma_\nu u_1) \times (\tilde{u}_4\gamma_\mu u_2)\mathbf{k}^{-2}\mathbf{q}^{-2}d^4k.$$

This integral on  $\mathbf{k}$  will be seen to be precisely the integral (12) for the radiative corrections to scattering, which we have worked out. The term may be combined with the renormalization terms resulting from the difference of the effects of mass change and the terms, Figs. 8f and 8g. Figures 8e, 8h, and 8i are similarly analyzed.

Finally the term Fig. 8c is clearly related to our vacuum polarization problem, and when integrated gives a term proportional to  $(\tilde{u}_4\gamma_\mu u_2)(\tilde{u}_3\gamma_\nu u_1)J_{\mu\nu}\mathbf{q}^{-4}$ . If the

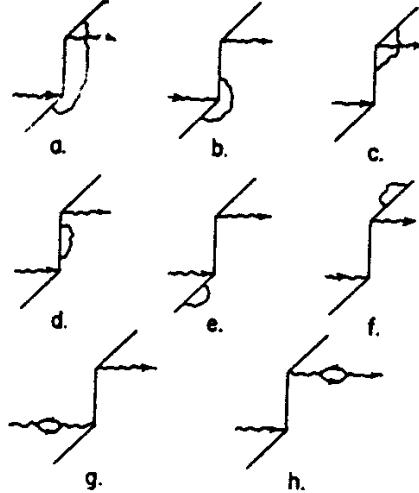


Figure 9: Radiative correction to the Compton scattering term (a) of Fig. 5. Appendix D.

charge is renormalized the term  $\ln(\lambda/m)$  in  $J_{\mu\nu}$  in (33) is omitted so there is no remaining dependence on the cut-off.

The only new integrals we require are the convergent integrals (38a) and (39a). They can be simplified by rationalizing the denominators and combining them by (14a). For example (38a) involves the factors  $(\mathbf{k}^2 - 2p_1 \cdot k)^{-1} (\mathbf{k}^2 + 2p_2 \cdot k)^{-1} \mathbf{k}^{-2} (\mathbf{q}^2 + \mathbf{k}^2 - 2q \cdot k)^{-2}$ . The first two may be combined by (14a) with a parameter  $x$ , and the second pair by an expression obtained by differentiation (l5a) with respect to  $b$  and calling the parameter  $y$ . There results a factor  $(\mathbf{k}^2 - 2p_x \cdot k)^{-2} (\mathbf{k}^2 + y\mathbf{q}^2 - 2yq \cdot k)^{-4}$  so that the integrals on  $d^4k$  now involve two factors and can be performed by the methods given earlier in the appendix. The subsequent integrals on the parameters  $x$  and  $y$  are complicated and have not been worked out in detail.

Working with charged mesons there is often a considerable reduction of the number of terms. For example, for the interaction between protons resulting from the exchange of two mesons only the term corresponding to Fig. 8h remains. Term 8a, for example, is impossible, for if the first proton emits a positive meson the second cannot absorb it directly for only neutrons can absorb positive mesons.

As a second example, consider the radiative correction to the Compton scattering. As seen from Eq. (15) and Fig. 5 this scattering is represented by two terms, so that we can consider the corrections to each one separately. Figure 9 shows the types of terms arising from corrections to the term of Fig. 5a. Calling  $\mathbf{k}$

the momentum of the virtual quantum, Fig. 9a gives an integral

$$\int \gamma_\mu(\mathbf{p}_2 - \mathbf{k} - m)^{-1} \mathbf{e}_2(\mathbf{p}_1 + \mathbf{q}_1 - \mathbf{k} - m)^{-1} \mathbf{e}_1(\mathbf{p}_1 - \mathbf{k} - m)^{-1} \gamma_\mu \mathbf{k}^{-2} d^4 k,$$

convergent without cut-off and reducible by the methods outlined in this appendix.

The other terms are relatively easy to evaluate. Terms  $b$  and  $c$  of Fig. 9 are closely related to radiative corrections (although somewhat more difficult to evaluate, for one of the states is not that of a free electron,  $(\mathbf{p}_1 + \mathbf{q})^2 \neq m^2$ ). Terms  $e, f$ , are renormalization terms. From term  $d$  must be subtracted explicitly the effect of mass  $\Delta m$ , as analyzed in Eqs. (26) and (27) leading to (28) with  $p' = \mathbf{p}_1 + \mathbf{q}$ ,  $\mathbf{a} = \mathbf{e}_2$ ,  $\mathbf{b} = \mathbf{e}_1$ . Terms  $g, h$  give zero since the vacuum polarization has zero effect on free light quanta,  $\mathbf{q}_1^2 = 0$ ,  $\mathbf{q}_2^2 = 0$ . The total is insensitive to the cut-off  $\lambda$ .

The result shows an infra-red catastrophe, the largest part of the effect. When cut-off at  $\lambda_{\min}$ , the effect proportional to  $\ln(m/\lambda_{\min})$  goes as

$$(e^2/\pi) \ln(m/\lambda_{\min})(1 - 2\theta \operatorname{ctn}2\theta), \quad (40a)$$

times the uncorrected amplitude, where  $(\mathbf{p}_2 - \mathbf{p}_1)^2 = 4m^2 \sin^2 \theta$ . This is the same as for the radiative correction to scattering for a deflection  $\mathbf{p}_2 - \mathbf{p}_1$ . This is physically clear since the long wave quanta are not effected by short-lived intermediate states. The infra-red effects arise<sup>30</sup> from a final adjustment of the field from the asymptotic coulomb field characteristic of the electron of momentum  $\mathbf{p}_1$  before the collision to that characteristic of an electron moving in a new direction  $\mathbf{p}_2$  after the collision.

The complete expression for the correction is a very complicated expression involving transcendental integrals.

As a final example we consider the interaction of a neutron with an electromagnetic field in virtue of the fact that the neutron may emit a virtual negative meson. We choose the example of pseudoscalar mesons with pseudovector coupling. The change in amplitude due to an electromagnetic field  $\mathbf{A} = \mathbf{a} \exp(-iq \cdot x)$  determines the scattering of a neutron by such a field. In the limit of small  $\mathbf{q}$  it will vary as  $\mathbf{q}\mathbf{a} - \mathbf{a}\mathbf{q}$  which represents the interaction of a particle possessing a magnetic moment. The first-order interaction between an electron and a neutron is given by the same calculation by considering the exchange of a quantum between the electron and the nucleon. In this case  $a_\mu$  is  $\mathbf{q}^{-2}$  times the matrix element of  $\gamma_\mu$  between the initial and final states of the electron, the states differing in momentum by  $\mathbf{q}$ .

The interaction may occur because the neutron of momentum  $\mathbf{p}_1$  emits a negative meson becoming a proton which proton interacts with the field and then reabsorbs the meson (Fig. 10a). The matrix for this process is  $(\mathbf{p}_2 = \mathbf{p}_1 + \mathbf{q})$ ,

$$\int (\gamma_5 \mathbf{k}(\mathbf{p}_2 - \mathbf{k} - M)^{-1} \mathbf{a}(\mathbf{p}_1 - \mathbf{k} - M)^{-1} (\gamma_5 \mathbf{k})(\mathbf{k}^2 - \mu^2)^{-1} d^4 k. \quad (41a)$$

Alternatively it may be the meson which interacts with the field. We assume that it does this in the manner of a scalar potential satisfying the Klein Gordon Eq.

---

<sup>30</sup>F. Bloch and A. Nordsieck, Phys. Rev. **52**, 54 (1937).

(35), (Fig. 10b)

$$\begin{aligned} & - \int (\gamma_5 \mathbf{k}_2) (\mathbf{p}_1 - \mathbf{k}_1 - M)^{-1} (\gamma_5 \mathbf{k}_1) (\mathbf{k}_2^2 - \mu^2)^{-1} \\ & \times (k_2 \cdot a + k_1 \cdot a) (\mathbf{k}_1^2 - \mu^2)^{-1} d^4 k_1, \end{aligned} \quad (42a)$$

where we have put  $\mathbf{k}_2 = \mathbf{k}_1 + \mathbf{q}$ . The change in sign arises because the virtual meson is negative. Finally there are two terms arising from the  $\gamma_5 \mathbf{a}$  part of the pseudovector coupling (Figs. 10c, 10d)

$$\int (\gamma_5 \mathbf{k}) (\mathbf{p}_2 - \mathbf{k} - M)^{-1} (\gamma_5 \mathbf{a}) (\mathbf{k}^2 - \mu^2)^{-1} d^4 k, \quad (43a)$$

and

$$\int (\gamma_5 \mathbf{a}) (\mathbf{p}_1 - \mathbf{k} - M)^{-1} (\gamma_5 \mathbf{k}) (\mathbf{k}^2 - \mu^2)^{-1} d^4 k, \quad (44a)$$

Using convergence factors in the manner discussed in the section on meson theories each integral can be evaluated and the results combined. Expanded in powers of  $\mathbf{q}$  the first term gives the magnetic moment of the neutron and is insensitive to the cut-off, the next gives the scattering amplitude of slow electrons on neutrons, and depends logarithmically on the cut-off.

The expressions may be simplified and combined somewhat before integration. This makes the integrals a little easier and also shows the relation to the case of pseudoscalar coupling. For example in (41a) the final  $\gamma_5 \mathbf{k}$  can be written as  $\gamma_5(\mathbf{k} - \mathbf{p}_1 + M)$  since  $\mathbf{p}_1 = M$  when operating on the initial neutron state. This is  $(\mathbf{p}_1 - \mathbf{k} - M)\gamma_5 + 2m\gamma_5$  since  $\gamma_5$  anticommutes with  $\mathbf{p}_1$  and  $\mathbf{k}$ . The first term cancels the  $(\mathbf{p}_1 - \mathbf{k} - M)^{-1}$  and gives a term which just cancels (43a). In a like manner the leading factor  $\gamma_5 \mathbf{k}$  in (41a) is written as  $-2M\gamma_5 - \gamma_5(\mathbf{p}_2 - \mathbf{k} - M)$ , the second term leading to a simpler term containing no  $(\mathbf{p}_2 - \mathbf{k} - M)^{-1}$  factor and combining with a similar one from (44a). One simplifies the  $\gamma_5 \mathbf{k}_1$  and  $\gamma_5 \mathbf{k}_2$  in (42a) in an analogous way. There finally results terms like (41a), (42a) but with pseudoscalar coupling  $2M\gamma_5$  instead of  $\gamma_5 \mathbf{k}$ , no terms like (43a) or (44a) and a remainder, representing the difference in effects of pseudovector and pseudoscalar coupling. The pseudoscalar terms do not depend sensitively on the cut-off, but the difference term depends on it logarithmically. The difference term affects the electron-neutron interaction but not the magnetic moment of the neutron.

Interaction of a proton with an electromagnetic potential can be similarly analyzed. There is an effect of virtual mesons on the electromagnetic properties of the proton even in the case that the mesons are neutral. It is analogous to the radiative corrections to the scattering of electrons due to virtual photons. The sum of the magnetic moments of neutron and proton for charged mesons is the same as the proton moment calculated for the corresponding neutral mesons. In fact it is readily seen by comparing diagrams. that for arbitrary  $\mathbf{q}$ , the scattering matrix to *first order in the electromagnetic potential* for a proton according to neutral meson theory is equal, if the mesons were charged, to the sum of the matrix for a neutron

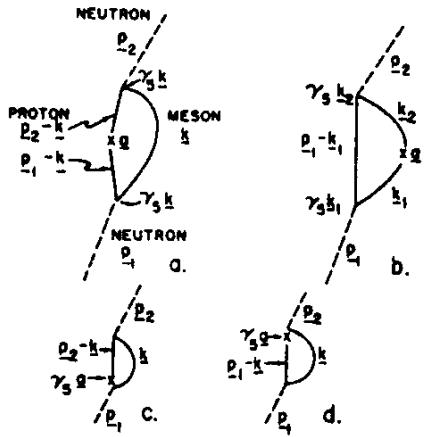


Figure 10: According to the meson theory a neutron interacts with an electromagnetic potential **a** by first emitting a virtual charged meson. The figure illustrates the case for a pseudoscalar meson with pseudovector coupling. Appendix D.

and the matrix for a proton. This is true, for any type or mixtures of meson coupling, to all orders in the coupling (neglecting the mass difference of neutron and proton).



## The Theory of Positrons

R. P. FEYNMAN

*Department of Physics, Cornell University, Ithaca, New York*

(Received April 8, 1949)

The problem of the behavior of positrons and electrons in given external potentials, neglecting their mutual interaction, is analyzed by replacing the theory of holes by a reinterpretation of the solutions of the Dirac equation. It is possible to write down a complete solution of the problem in terms of boundary conditions on the wave function, and this solution contains automatically all the possibilities of virtual (and real) pair formation and annihilation together with the ordinary scattering processes, including the correct relative signs of the various terms.

In this solution, the "negative energy states" appear in a form which may be pictured (as by Stückelberg) in space-time as waves traveling away from the external potential backwards in time. Experimentally, such a wave corresponds to a positron approaching the potential and annihilating the electron. A particle moving forward in time (electron) in a potential may be scattered forward in time (ordinary scattering) or backward (pair annihilation). When moving backward (positron) it may be scattered backward

in time (positron scattering) or forward (pair production). For such a particle the amplitude for transition from an initial to a final state is analyzed to any order in the potential by considering it to undergo a sequence of such scatterings.

The amplitude for a process involving many such particles is the product of the transition amplitudes for each particle. The exclusion principle requires that antisymmetric combinations of amplitudes be chosen for those complete processes which differ only by exchange of particles. It seems that a consistent interpretation is only possible if the exclusion principle is adopted. The exclusion principle need not be taken into account in intermediate states. Vacuum problems do not arise for charges which do not interact with one another, but these are analyzed nevertheless in anticipation of application to quantum electrodynamics.

The results are also expressed in momentum-energy variables. Equivalence to the second quantization theory of holes is proved in an appendix.

### 1. INTRODUCTION

THIS is the first of a set of papers dealing with the solution of problems in quantum electrodynamics. The main principle is to deal directly with the solutions to the Hamiltonian differential equations rather than with these equations themselves. Here we treat simply the motion of electrons and positrons in given external potentials. In a second paper we consider the interactions of these particles, that is, quantum electrodynamics.

The problem of charges in a fixed potential is usually treated by the method of second quantization of the electron field, using the ideas of the theory of holes. Instead we show that by a suitable choice and interpretation of the solutions of Dirac's equation the problem may be equally well treated in a manner which is fundamentally no more complicated than Schrödinger's method of dealing with one or more particles. The various creation and annihilation operators in the conventional electron field view are required because the number of particles is not conserved, i.e., pairs may be created or destroyed. On the other hand charge is conserved which suggests that if we follow the charge, not the particle, the results can be simplified.

In the approximation of classical relativistic theory the creation of an electron pair (electron *A*, positron *B*) might be represented by the start of two world lines from the point of creation, 1. The world lines of the positron will then continue until it annihilates another electron, *C*, at a world point 2. Between the times  $t_1$  and  $t_2$  there are then three world lines, before and after only one. However, the world lines of *C*, *B*, and *A* together form one continuous line albeit the "positron part" *B* of this continuous line is directed backwards in time. Following the charge rather than the particles corresponds to considering this continuous world line

as a whole rather than breaking it up into its pieces. It is as though a bombardier flying low over a road suddenly sees three roads and it is only when two of them come together and disappear again that he realizes that he has simply passed over a long switchback in a single road.

This over-all space-time point of view leads to considerable simplification in many problems. One can take into account at the same time processes which ordinarily would have to be considered separately. For example, when considering the scattering of an electron by a potential one automatically takes into account the effects of virtual pair productions. The same equation, Dirac's, which describes the deflection of the world line of an electron in a field, can also describe the deflection (and in just as simple a manner) when it is large enough to reverse the time-sense of the world line, and thereby correspond to pair annihilation. Quantum mechanically the direction of the world lines is replaced by the direction of propagation of waves.

This view is quite different from that of the Hamiltonian method which considers the future as developing continuously from out of the past. Here we imagine the entire space-time history laid out, and that we just become aware of increasing portions of it successively. In a scattering problem this over-all view of the complete scattering process is similar to the *S*-matrix viewpoint of Heisenberg. The temporal order of events during the scattering, which is analyzed in such detail by the Hamiltonian differential equation, is irrelevant. The relation of these viewpoints will be discussed much more fully in the introduction to the second paper, in which the more complicated interactions are analyzed.

The development stemmed from the idea that in non-relativistic quantum mechanics the amplitude for a given process can be considered as the sum of an ampli-

tude for each space-time path available.<sup>1</sup> In view of the fact that in classical physics positrons could be viewed as electrons proceeding along world lines toward the past (reference 7) the attempt was made to remove, in the relativistic case, the restriction that the paths must proceed always in one direction in time. It was discovered that the results could be even more easily understood from a more familiar physical viewpoint, that of scattered waves. This viewpoint is the one used in this paper. After the equations were worked out physically the proof of the equivalence to the second quantization theory was found.<sup>2</sup>

First we discuss the relation of the Hamiltonian differential equation to its solution, using for an example the Schrödinger equation. Next we deal in an analogous way with the Dirac equation and show how the solutions may be interpreted to apply to positrons. The interpretation seems not to be consistent unless the electrons obey the exclusion principle. (Charges obeying the Klein-Gordon equations can be described in an analogous manner, but here consistency apparently requires Bose statistics.)<sup>3</sup> A representation in momentum and energy variables which is useful for the calculation of matrix elements is described. A proof of the equivalence of the method to the theory of holes in second quantization is given in the Appendix.

## 2. GREEN'S FUNCTION TREATMENT OF SCHRÖDINGER'S EQUATION

We begin by a brief discussion of the relation of the non-relativistic wave equation to its solution. The ideas will then be extended to relativistic particles, satisfying Dirac's equation, and finally in the succeeding paper to interacting relativistic particles, that is, quantum electrodynamics.

The Schrödinger equation

$$i\partial\psi/\partial t = H\psi, \quad (1)$$

describes the change in the wave function  $\psi$  in an infinitesimal time  $\Delta t$  as due to the operation of an operator  $\exp(-iH\Delta t)$ . One can ask also, if  $\psi(\mathbf{x}_1, t_1)$  is the wave function at  $\mathbf{x}_1$  at time  $t_1$ , what is the wave function at time  $t_2 > t_1$ ? It can always be written as

$$\psi(\mathbf{x}_2, t_2) = \int K(\mathbf{x}_2, t_2; \mathbf{x}_1, t_1) \psi(\mathbf{x}_1, t_1) d^3\mathbf{x}_1, \quad (2)$$

where  $K$  is a Green's function for the linear Eq. (1). (We have limited ourselves to a single particle of coordinate  $\mathbf{x}$ , but the equations are obviously of greater generality.) If  $H$  is a constant operator having eigenvalues  $E_n$ , eigenfunctions  $\phi_n$  so that  $\psi(\mathbf{x}, t_1)$  can be expanded as  $\sum_n C_n \phi_n(\mathbf{x})$ , then  $\psi(\mathbf{x}, t_2) = \exp(-iE_n(t_2 - t_1)) \times C_n \phi_n(\mathbf{x})$ . Since  $C_n = \int \phi_n^*(\mathbf{x}_1) \psi(\mathbf{x}_1, t_1) d^3\mathbf{x}_1$ , one finds

<sup>1</sup> R. P. Feynman, Rev. Mod. Phys. **20**, 367 (1948).

<sup>2</sup> The equivalence of the entire procedure (including photon interactions) with the work of Schwinger and Tomonaga has been demonstrated by F. J. Dyson, Phys. Rev. **75**, 486 (1949).

<sup>3</sup> These are special examples of the general relation of spin and statistics deduced by W. Pauli, Phys. Rev. **58**, 716 (1940).

(where we write 1 for  $\mathbf{x}_1, t_1$  and 2 for  $\mathbf{x}_2, t_2$ ) in this case

$$K(2, 1) = \sum_n \phi_n(\mathbf{x}_2) \phi_n^*(\mathbf{x}_1) \exp(-iE_n(t_2 - t_1)), \quad (3)$$

for  $t_2 > t_1$ . We shall find it convenient for  $t_2 < t_1$  to define  $K(2, 1) = 0$  (Eq. (2) is then not valid for  $t_2 < t_1$ ). It is then readily shown that in general  $K$  can be defined by that solution of

$$(i\partial/\partial t_2 - H_2) K(2, 1) = i\delta(2, 1), \quad (4)$$

which is zero for  $t_2 < t_1$ , where  $\delta(2, 1) = \delta(t_2 - t_1)\delta(\mathbf{x}_2 - \mathbf{x}_1) \times \delta(y_2 - y_1)\delta(z_2 - z_1)$  and the subscript 2 on  $H_2$  means that the operator acts on the variables of 2 of  $K(2, 1)$ . When  $H$  is not constant, (2) and (4) are valid but  $K$  is less easy to evaluate than (3).<sup>4</sup>

We can call  $K(2, 1)$  the total amplitude for arrival at  $\mathbf{x}_2, t_2$  starting from  $\mathbf{x}_1, t_1$ . (It results from adding an amplitude,  $\exp(iS)$ , for each space time path between these points, where  $S$  is the action along the path!) The transition amplitude for finding a particle in state  $\chi(\mathbf{x}_2, t_2)$  at time  $t_2$ , if at  $t_1$  it was in  $\psi(\mathbf{x}_1, t_1)$ , is

$$\int \chi^*(2) K(2, 1) \psi(1) d^3\mathbf{x}_1 d^3\mathbf{x}_2. \quad (5)$$

A quantum mechanical system is described equally well by specifying the function  $K$ , or by specifying the Hamiltonian  $H$  from which it results. For some purposes the specification in terms of  $K$  is easier to use and visualize. We desire eventually to discuss quantum electrodynamics from this point of view.

To gain a greater familiarity with the  $K$  function and the point of view it suggests, we consider a simple perturbation problem. Imagine we have a particle in a weak potential  $U(\mathbf{x}, t)$ , a function of position and time. We wish to calculate  $K(2, 1)$  if  $U$  differs from zero only for  $t$  between  $t_1$  and  $t_2$ . We shall expand  $K$  in increasing powers of  $U$ :

$$K(2, 1) = K_0(2, 1) + K^{(1)}(2, 1) + K^{(2)}(2, 1) + \dots \quad (6)$$

To zero order in  $U$ ,  $K$  is that for a free particle,  $K_0(2, 1)$ .<sup>4</sup> To study the first order correction  $K^{(1)}(2, 1)$ , first consider the case that  $U$  differs from zero only for the infinitesimal time interval  $\Delta t_3$  between some time  $t_3$  and  $t_3 + \Delta t_3$  ( $t_1 < t_3 < t_2$ ). Then if  $\psi(1)$  is the wave function at  $\mathbf{x}_1, t_1$ , the wave function at  $\mathbf{x}_3, t_3$  is

$$\psi(3) = \int K_0(3, 1) \psi(1) d^3\mathbf{x}_1, \quad (7)$$

since from  $t_1$  to  $t_3$  the particle is free. For the short interval  $\Delta t_3$  we solve (1) as

$$\begin{aligned} \psi(\mathbf{x}, t_3 + \Delta t_3) &= \exp(-iH\Delta t_3) \psi(\mathbf{x}, t_3) \\ &= (1 - iH_0\Delta t_3 - iU\Delta t_3) \psi(\mathbf{x}, t_3), \end{aligned}$$

<sup>4</sup> For a non-relativistic free particle, where  $\phi_n = \exp(ip \cdot x)$ ,  $E_n = p^2/2m$ , (3) gives, as is well known

$$\begin{aligned} K_0(2, 1) &= \int \exp[-(ip \cdot \mathbf{x}_1 - ip \cdot \mathbf{x}_2) - ip^2(t_2 - t_1)/2m] d^3\mathbf{p} (2\pi)^{-3} \\ &= (2\pi m^{-1}(t_2 - t_1))^{-1} \exp(\frac{1}{2}im(\mathbf{x}_2 - \mathbf{x}_1)^2(t_2 - t_1)^{-1}) \end{aligned}$$

for  $t_2 > t_1$ , and  $K_0 = 0$  for  $t_2 < t_1$ .

where we put  $H = H_0 + U$ ,  $H_0$  being the Hamiltonian of a free particle. Thus  $\psi(\mathbf{x}, t_3 + \Delta t_3)$  differs from what it would be if the potential were zero (namely  $(1 - iH_0\Delta t_3)\psi(\mathbf{x}, t_3)$ ) by the extra piece

$$\Delta\psi = -iU(\mathbf{x}_3, t_3)\cdot\psi(\mathbf{x}_3, t_3)\Delta t_3, \quad (8)$$

which we shall call the amplitude scattered by the potential. The wave function at 2 is given by

$$\psi(\mathbf{x}_2, t_2) = \int K_0(\mathbf{x}_2, t_2; \mathbf{x}_3, t_3 + \Delta t_3)\psi(\mathbf{x}_3, t_3 + \Delta t_3)d^3\mathbf{x}_3, \quad (9)$$

since after  $t_3 + \Delta t_3$  the particle is again free. Therefore the change in the wave function at 2 brought about by the potential is (substitute (7) into (8) and (8) into the equation for  $\psi(\mathbf{x}_2, t_2)$ ):

$$\Delta\psi(2) = -i \int K_0(2, 3)U(3)K_0(3, 1)\psi(1)d^3\mathbf{x}_1d^3\mathbf{x}_3\Delta t_3.$$

In the case that the potential exists for an extended time, it may be looked upon as a sum of effects from each interval  $\Delta t_3$  so that the total effect is obtained by integrating over  $t_3$  as well as  $\mathbf{x}_3$ . From the definition (2) of  $K$  then, we find

$$K^{(1)}(2, 1) = -i \int K_0(2, 3)U(3)K_0(3, 1)d\tau_3, \quad (9)$$

where the integral can now be extended over all space and time,  $d\tau_3 = d^3\mathbf{x}_3 dt_3$ . Automatically there will be no contribution if  $t_3$  is outside the range  $t_1$  to  $t_2$  because of our definition,  $K_0(2, 1) = 0$  for  $t_2 < t_1$ .

We can understand the result (6), (9) this way. We can imagine that a particle travels as a free particle from point to point, but is scattered by the potential  $U$ . Thus the total amplitude for arrival at 2 from 1 can be considered as the sum of the amplitudes for various alternative routes. It may go directly from 1 to 2 (amplitude  $K_0(2, 1)$ , giving the zero order term in (6)). Or (see Fig. 1(a)) it may go from 1 to 3 (amplitude  $K_0(3, 1)$ ), get scattered there by the potential (scattering amplitude  $-iU(3)$  per unit volume and time) and then go from 3 to 2 (amplitude  $K_0(2, 3)$ ). This may occur for any point 3 so that summing over these alternatives gives (9).

Again, it may be scattered twice by the potential (Fig. 1(b)). It goes from 1 to 3 ( $K_0(3, 1)$ ), gets scattered there ( $-iU(3)$ ) then proceeds to some other point, 4, in space time (amplitude  $K_0(4, 3)$ ) is scattered again ( $-iU(4)$ ) and then proceeds to 2 ( $K_0(2, 4)$ ). Summing over all possible places and times for 3, 4 find that the second order contribution to the total amplitude  $K^{(2)}(2, 1)$  is

$$(-i)^2 \int \int K_0(2, 4)U(4)K_0(4, 3) \times U(3)K_0(3, 1)d\tau_3 d\tau_4. \quad (10)$$

This can be readily verified directly from (1) just as (9)

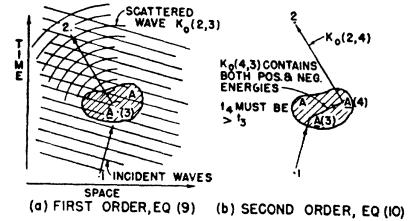


FIG. 1. The Schrödinger (and Dirac) equation can be visualized as describing the fact that plane waves are scattered successively by a potential. Figure 1 (a) illustrates the situation in first order.  $K_0(2, 3)$  is the amplitude for a free particle starting at point 3 to arrive at 2. The shaded region indicates the presence of the potential  $A$  which scatters at 3 with amplitude  $-iA(3)$  per  $\text{cm}^3\text{sec}$ . (Eq. (9)). In (b) is illustrated the second order process (Eq. (10)), the waves scattered at 3 are scattered again at 4. However, in Dirac one-electron theory  $K_0(4, 3)$  would represent electrons both of positive and of negative energies proceeding from 3 to 4. This is remedied by choosing a different scattering kernel  $K_+(4, 3)$ , Fig. 2.

was. One can in this way obviously write down any of the terms of the expansion (6).<sup>5</sup>

### 3. TREATMENT OF THE DIRAC EQUATION

We shall now extend the method of the last section to apply to the Dirac equation. All that would seem to be necessary in the previous equations is to consider  $H$  as the Dirac Hamiltonian,  $\psi$  as a symbol with four indices (for each particle). Then  $K_0$  can still be defined by (3) or (4) and is now a 4-4 matrix which operating on the initial wave function, gives the final wave function. In (10),  $U(3)$  can be generalized to  $A_4(3) - \alpha \cdot A(3)$  where  $A_4, A$  are the scalar and vector potential (times  $e$ ), the electron charge) and  $\alpha$  are Dirac matrices.

To discuss this we shall define a convenient relativistic notation. We represent four-vectors like  $\mathbf{x}, t$  by a symbol  $x_\mu$ , where  $\mu = 1, 2, 3, 4$  and  $x_4 = t$  is real. Thus the vector and scalar potential (times  $e$ )  $A, A_4$  is  $A_\mu$ . The four matrices  $\beta\alpha, \beta$  can be considered as transforming as a four vector  $\gamma_\mu$  (our  $\gamma_\mu$  differs from Pauli's by a factor  $i$  for  $\mu = 1, 2, 3$ ). We use the summation convention  $a_\mu b_\mu = a_4 b_4 - a_1 b_1 - a_2 b_2 - a_3 b_3 = a \cdot b$ . In particular if  $a_\mu$  is any four vector (but not a matrix) we write  $a = a_\mu \gamma_\mu$  so that  $a$  is a matrix associated with a vector ( $a$  will often be used in place of  $a_\mu$  as a symbol for the vector). The  $\gamma_\mu$  satisfy  $\gamma_\mu \gamma_\nu + \gamma_\nu \gamma_\mu = 2\delta_{\mu\nu}$  where  $\delta_{44} = +1$ ,  $\delta_{11} = \delta_{22} = \delta_{33} = -1$ , and the other  $\delta_{\mu\nu}$  are zero. As a consequence of our summation convention  $\delta_{\mu\nu} a_\nu = a_\mu$  and  $\delta_{\mu\mu} = 4$ . Note that  $a b + b a = 2a \cdot b$  and that  $a^\mu = a_\mu a^\mu = a \cdot a$  is a pure number. The symbol  $\partial/\partial x_\mu$  will mean  $\partial/\partial t$  for  $\mu = 4$ , and  $-\partial/\partial x_1, -\partial/\partial y, -\partial/\partial z$  for  $\mu = 1, 2, 3$ . Call  $\nabla = \gamma_\mu \partial/\partial x_\mu = \beta \partial/\partial t + \beta \alpha \cdot \nabla$ . We shall imagine

<sup>5</sup> We are simply solving by successive approximations an integral equation (deducible directly from (1) with  $H = H_0 + U$  and (4) with  $H = H_0$ ),

$$\psi(2) = -i \int K_0(2, 3)U(3)\psi(3)d\tau_3 + \int K_0(2, 1)\psi(1)d^3\mathbf{x}_1,$$

where the first integral extends over all space and all times  $t_3$  greater than the  $t_1$  appearing in the second term, and  $t_2 > t_1$ .

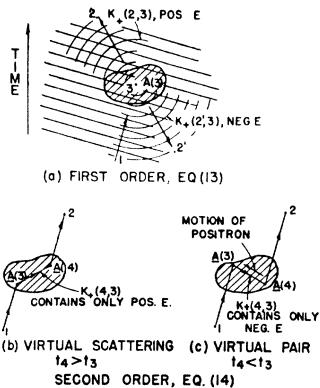


FIG. 2. The Dirac equation permits another solution  $K_+(2, 1)$  if one considers that waves scattered by the potential can proceed backwards in time as in Fig. 2 (a). This is interpreted in the second order processes (b), (c), by noting that there is now the possibility (c) of virtual pair production at 4, the positron going to 3 to be annihilated. This can be pictured as similar to ordinary scattering (b) except that the electron is scattered backwards in time from 3 to 4. The waves scattered from 3 to 2' in (a) represent the possibility of a positron arriving at 3 from 2' and annihilating the electron from 1. This view is proved equivalent to hole theory: electrons traveling backwards in time are recognized as positrons.

hereafter, purely for relativistic convenience, that  $\phi_n^*$  in (3) is replaced by its adjoint  $\bar{\phi}_n = \phi_n^* \beta$ .

Thus the Dirac equation for a particle, mass  $m$ , in an external field  $A = A_\mu \gamma^\mu$  is

$$(i\nabla - m)\psi = A\psi, \quad (11)$$

and Eq. (4) determining the propagation of a free particle becomes

$$(i\nabla_2 - m)K_+(2, 1) = i\delta(2, 1), \quad (12)$$

the index 2 on  $\nabla_2$  indicating differentiation with respect to the coordinates  $x_{2\mu}$  which are represented as 2 in  $K_+(2, 1)$  and  $\delta(2, 1)$ .

The function  $K_+(2, 1)$  is defined in the absence of a field. If a potential  $A$  is acting a similar function, say  $K_+^{(A)}(2, 1)$  can be defined. It differs from  $K_+(2, 1)$  by a first order correction given by the analogue of (9) namely

$$K_+^{(1)}(2, 1) = -i \int K_+(2, 3)A(3)K_+(3, 1)d\tau_3, \quad (13)$$

representing the amplitude to go from 1 to 3 as a free particle, get scattered there by the potential (now the matrix  $A(3)$  instead of  $U(3)$ ) and continue to 2 as free. The second order correction, analogous to (10) is

$$K_+^{(2)}(2, 1) = - \int \int K_+(2, 4)A(4) \times K_+(4, 3)A(3)K_+(3, 1)d\tau_4 d\tau_3, \quad (14)$$

and so on. In general  $K_+^{(A)}$  satisfies

$$(i\nabla_2 - A(2) - m)K_+^{(A)}(2, 1) = i\delta(2, 1), \quad (15)$$

and the successive terms (13), (14) are the power series

expansion of the integral equation

$$K_+^{(A)}(2, 1) = K_+(2, 1)$$

$$-i \int K_+(2, 3)A(3)K_+(3, 1)d\tau_3, \quad (16)$$

which it also satisfies.

We would now expect to choose, for the special solution of (12),  $K_+ = K_0$  where  $K_0(2, 1)$  vanishes for  $t_2 < t_1$  and for  $t_2 > t_1$  is given by (3) where  $\phi_n$  and  $E_n$  are the eigenfunctions and energy values of a particle satisfying Dirac's equation, and  $\phi_n^*$  is replaced by  $\bar{\phi}_n$ .

The formulas arising from this choice, however, suffer from the drawback that they apply to the one electron theory of Dirac rather than to the hole theory of the positron. For example, consider as in Fig. 1(a) an electron after being scattered by a potential in a small region 3 of space time. The one electron theory says (as does (3) with  $K_+ = K_0$ ) that the scattered amplitude at another point 2 will proceed toward positive times with both positive and negative energies, that is with both positive and negative rates of change of phase. No wave is scattered to times previous to the time of scattering. These are just the properties of  $K_0(2, 3)$ .

On the other hand, according to the positron theory negative energy states are not available to the electron after the scattering. Therefore the choice  $K_+ = K_0$  is unsatisfactory. But there are other solutions of (12). We shall choose the solution defining  $K_+(2, 1)$  so that  $K_+(2, 1)$  for  $t_2 > t_1$  is the sum of (3) over positive energy states only. Now this new solution must satisfy (12) for all times in order that the representation be complete. It must therefore differ from the old solution  $K_0$  by a solution of the homogeneous Dirac equation. It is clear from the definition that the difference  $K_0 - K_+$  is the sum of (3) over all negative energy states, as long as  $t_2 > t_1$ . But this difference must be a solution of the homogeneous Dirac equation for all times and must therefore be represented by the same sum over negative energy states also for  $t_2 < t_1$ . Since  $K_0 = 0$  in this case, it follows that our new kernel,  $K_+(2, 1)$ , for  $t_2 < t_1$  is the negative of the sum (3) over negative energy states. That is,

$$\begin{aligned} K_+(2, 1) &= \sum_{POS E_n} \phi_n(2) \bar{\phi}_n(1) \\ &\quad \times \exp(-iE_n(t_2 - t_1)) \quad \text{for } t_2 > t_1 \\ &= -\sum_{NEG E_n} \phi_n(2) \bar{\phi}_n(1) \\ &\quad \times \exp(-iE_n(t_2 - t_1)) \quad \text{for } t_2 < t_1. \end{aligned} \quad (17)$$

With this choice of  $K_+$  our equations such as (13) and (14) will now give results equivalent to those of the positron hole theory.

That (14), for example, is the correct second order expression for finding at 2 an electron originally at 1 according to the positron theory may be seen as follows (Fig. 2). Assume as a special example that  $t_2 > t_1$  and that the potential vanishes except in interval  $t_2 - t_1$  so that  $t_4$  and  $t_3$  both lie between  $t_1$  and  $t_2$ .

First suppose  $t_4 > t_3$  (Fig. 2(b)). Then (since  $t_3 > t_1$ )

the electron assumed originally in a positive energy state propagates in that state (by  $K_+(3, 1)$ ) to position 3 where it gets scattered ( $A(3)$ ). It then proceeds to 4, which it must do as a positive energy electron. This is correctly described by (14) for  $K_+(4, 3)$  contains only positive energy components in its expansion, as  $t_4 > t_3$ . After being scattered at 4 it then proceeds on to 2, again necessarily in a positive energy state, as  $t_2 > t_4$ .

In positron theory there is an additional contribution due to the possibility of virtual pair production (Fig. 2(c)). A pair could be created by the potential  $A(4)$  at 4, the electron of which is that found later at 2. The positron (or rather, the hole) proceeds to 3 where it annihilates the electron which has arrived there from 1.

This alternative is already included in (14) as contributions for which  $t_4 < t_3$ , and its study will lead us to an interpretation of  $K_+(4, 3)$  for  $t_4 < t_3$ . The factor  $K_+(2, 4)$  describes the electron (after the pair production at 4) proceeding from 4 to 2. Likewise  $K_+(3, 1)$  represents the electron proceeding from 1 to 3.  $K_+(4, 3)$  must therefore represent the propagation of the positron or hole from 4 to 3. That it does so is clear. The fact that in hole theory the hole proceeds in the manner of an electron of negative energy is reflected in the fact that  $K_+(4, 3)$  for  $t_4 < t_3$  is (minus) the sum of only negative energy components. In hole theory the real energy of these intermediate states is, of course, positive. This is true here too, since in the phases  $\exp(-iE_n(t_4 - t_3))$  defining  $K_+(4, 3)$  in (17),  $E_n$  is negative but so is  $t_4 - t_3$ . That is, the contributions vary with  $t_3$  as  $\exp(-i|E_n|(t_3 - t_4))$  as they would if the energy of the intermediate state were  $|E_n|$ . The fact that the entire sum is taken as negative in computing  $K_+(4, 3)$  is reflected in the fact that in hole theory the amplitude has its sign reversed in accordance with the Pauli principle and the fact that the electron arriving at 2 has been exchanged with one in the sea.<sup>6</sup> To this, and to higher orders, all processes involving virtual pairs are correctly described in this way.

The expressions such as (14) can still be described as a passage of the electron from 1 to 3 ( $K_+(3, 1)$ ), scattering at 3 by  $A(3)$ , proceeding to 4 ( $K_+(4, 3)$ ), scattering again,  $A(4)$ , arriving finally at 2. The scatterings may, however, be toward both future and past times, an electron propagating backwards in time being recognized as a positron.

This therefore suggests that negative energy components created by scattering in a potential be considered as waves propagating from the scattering point toward the past, and that such waves represent the propagation of a positron annihilating the electron in the potential.<sup>7</sup>

<sup>6</sup> It has often been noted that the one-electron theory apparently gives the same matrix elements for this process as does hole theory. The problem is one of interpretation, especially in a way that will also give correct results for other processes, e.g., self-energy.

<sup>7</sup> The idea that positrons can be represented as electrons with proper time reversed relative to true time has been discussed by the author and others, particularly by Stückelberg. E. C. C.

With this interpretation real pair production is also described correctly (see Fig. 3). For example in (13) if  $t_1 < t_3 < t_2$  the equation gives the amplitude that if at time  $t_1$  one electron is present at 1, then at time  $t_2$  just one electron will be present (having been scattered at 3) and it will be at 2. On the other hand if  $t_2$  is less than  $t_3$ , for example, if  $t_2 = t_1 < t_3$ , the same expression gives the amplitude that a pair, electron at 1, positron at 2 will annihilate at 3, and subsequently no particles will be present. Likewise if  $t_2$  and  $t_1$  exceed  $t_3$  we have (minus) the amplitude for finding a single pair, electron at 2, positron at 1 created by  $A(3)$  from a vacuum. If  $t_1 > t_3 > t_2$ , (13) describes the scattering of a positron. All these amplitudes are relative to the amplitude that a vacuum will remain a vacuum, which is taken as unity. (This will be discussed more fully later.)

The analogue of (2) can be easily worked out.<sup>8</sup> It is,

$$\psi(2) = \int K_+(2, 1)N(1)\psi(1)d^3V_1, \quad (18)$$

where  $d^3V_1$  is the volume element of the closed 3-dimensional surface of a region of space time containing

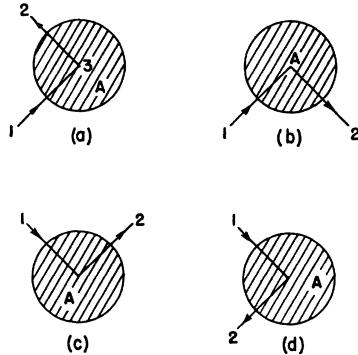


FIG. 3. Several different processes can be described by the same formula depending on the time relations of the variables  $t_2$ ,  $t_1$ . Thus  $P_s |K_+^{(A)}(2, 1)|^2$  is the probability that: (a) An electron at 1 will be scattered at 2 (and no other pairs form in vacuum). (b) Electron at 1 and positron at 2 annihilate leaving nothing. (c) A single pair at 1 and 2 is created from vacuum. (d) A positron at 2 is scattered to 1.  $(K_+^{(A)}(2, 1))$  is the sum of the effects of scattering in the potential to all orders.  $P_s$  is a normalizing constant.)

Stückelberg, Helv. Phys. Acta 15, 23 (1942); R. P. Feynman, Phys. Rev. 74, 939 (1948). The fact that classically the action (proper time) increases continuously as one follows a trajectory is reflected in quantum mechanics in the fact that the phase, which is  $|E_n| |t_2 - t_1|$ , always increases as the particle proceeds from one scattering point to the next.

<sup>8</sup> By multiplying (12) on the right by  $(-\iota\nabla_1 - m)$  and noting that  $\nabla_1\delta(2, 1) = -\nabla_2\delta(2, 1)$  show that  $K_+(2, 1)$  also satisfies  $K_+(2, 1)(-\iota\nabla_1 - m) = i\delta(2, 1)$ , where the  $\nabla_1$  operates on variable 1 in  $K_+(2, 1)$  but is written after that function to keep the correct order of the  $\gamma$  matrices. Multiply this equation by  $\psi(1)$  and Eq. (11) (with  $A=0$ , calling the variables 1) by  $K_+(2, 1)$ , subtract and integrate over a region of space-time. The integral on the left-hand side can be transformed to an integral over the surface of the region. The right-hand side is  $\psi(2)$  if the point 2 lies within the region, and is zero otherwise. (What happens when the 3-surface contains a light line and hence has no unique normal need not concern us as these points can be made to occur so far away from 2 that their contribution vanishes.)

point 2, and  $N(1)$  is  $N_\mu(1)\gamma_\mu$  where  $N_\mu(1)$  is the *inward* drawn unit normal to the surface at the point 1. That is, the wave function  $\psi(2)$  (in this case for a free particle) is determined at any point inside a four-dimensional region if its values on the surface of that region are specified.

To interpret this, consider the case that the 3-surface consists essentially of all space at some time say  $t=0$  previous to  $t_2$ , and of all space at the time  $T>t_2$ . The cylinder connecting these to complete the closure of the surface may be very distant from  $\mathbf{x}_2$  so that it gives no appreciable contribution (as  $K_+(2, 1)$  decreases exponentially in space-like directions). Hence, if  $\gamma_4=\beta$ , since the inward drawn normals  $N$  will be  $\beta$  and  $-\beta$ ,

$$\begin{aligned}\psi(2) = & \int K_+(2, 1)\beta\psi(1)d^3\mathbf{x}_1 \\ & - \int K_+(2, 1')\beta\psi(1')d^3\mathbf{x}_{1'},\end{aligned}\quad (19)$$

where  $t_1=0$ ,  $t_{1'}=T$ . Only positive energy (electron) components in  $\psi(1)$  contribute to the first integral and only negative energy (positron) components of  $\psi(1')$  to the second. That is, the amplitude for finding a charge at 2 is determined both by the amplitude for finding an electron previous to the measurement and by the amplitude for finding a positron after the measurement. This might be interpreted as meaning that even in a problem involving but one charge the amplitude for finding the charge at 2 is not determined when the only thing known in the amplitude for finding an electron (or a positron) at an earlier time. There may have been no electron present initially but a pair was created in the measurement (or also by other external fields). The amplitude for this contingency is specified by the amplitude for finding a positron in the future.

We can also obtain expressions for transition amplitudes, like (5). For example if at  $t=0$  we have an electron present in a state with (positive energy) wave function  $f(\mathbf{x})$ , what is the amplitude for finding it at  $t=T$  with the (positive energy) wave function  $g(\mathbf{x})$ ? The amplitude for finding the electron anywhere after  $t=0$  is given by (19) with  $\psi(1)$  replaced by  $f(\mathbf{x})$ , the second integral vanishing. Hence, the transition element to find it in state  $g(\mathbf{x})$  is, in analogy to (5), just ( $t_2=T$ ,  $t_1=0$ )

$$\int \bar{g}(\mathbf{x}_2)\beta K_+(2, 1)\beta f(\mathbf{x}_1)d^3\mathbf{x}_1d^3\mathbf{x}_2,\quad (20)$$

since  $g^*=\bar{g}\beta$ .

If a potential acts somewhere in the interval between 0 and  $T$ ,  $K_+$  is replaced by  $K_+^{(4)}$ . Thus the first order effect on the transition amplitude is, from (13),

$$-i \int \bar{g}(\mathbf{x}_2)\beta K_+(2, 3)\mathbf{A}(3)K_+(3, 1)\beta f(\mathbf{x}_1)d^3\mathbf{x}_1d^3\mathbf{x}_2.\quad (21)$$

Expressions such as this can be simplified and the 3-surface integrals, which are inconvenient for rela-

tivistic calculations, can be removed as follows. Instead of defining a state by the wave function  $f(\mathbf{x})$ , which it has at a given time  $t_1=0$ , we define the state by the function  $f(1)$  of four variables  $\mathbf{x}_1, t_1$  which is a solution of the free particle equation for all  $t_1$  and is  $f(\mathbf{x}_1)$  for  $t_1=0$ . The final state is likewise defined by a function  $g(2)$  over-all space-time. Then our surface integrals can be performed since  $\int K_+(3, 1)\beta f(\mathbf{x}_1)d^3\mathbf{x}_1=f(3)$  and  $\int \bar{g}(\mathbf{x}_2)\beta d^3\mathbf{x}_2 K_+(2, 3)=\bar{g}(3)$ . There results

$$-i \int \bar{g}(3)\mathbf{A}(3)f(3)d\tau_3,\quad (22)$$

the integral now being over-all space-time. The transition amplitude to second order (from (14)) is

$$- \int \int \bar{g}(2)\mathbf{A}(2)K_+(2, 1)\mathbf{A}(1)f(1)d\tau_1d\tau_2,\quad (23)$$

for the particle arriving at 1 with amplitude  $f(1)$  is scattered ( $\mathbf{A}(1)$ ), progresses to 2, ( $K_+(2, 1)$ ), and is scattered again ( $\mathbf{A}(2)$ ), and we then ask for the amplitude that it is in state  $g(2)$ . If  $g(2)$  is a negative energy state we are solving a problem of annihilation of electron in  $f(1)$ , positron in  $g(2)$ , etc.

We have been emphasizing scattering problems, but obviously the motion in a fixed potential  $V$ , say in a hydrogen atom, can also be dealt with. If it is first viewed as a scattering problem we can ask for the amplitude,  $\phi_k(1)$ , that an electron with original free wave function was scattered  $k$  times in the potential  $V$  either forward or backward in time to arrive at 1. Then the amplitude after one more scattering is

$$\phi_{k+1}(2) = -i \int K_+(2, 1)V(1)\phi_k(1)d\tau_1.\quad (24)$$

An equation for the total amplitude

$$\psi(1) = \sum_{k=0}^{\infty} \phi_k(1)$$

for arriving at 1 either directly or after any number of scatterings is obtained by summing (24) over all  $k$  from 0 to  $\infty$ ;

$$\psi(2) = \phi_0(2) - i \int K_+(2, 1)V(1)\psi(1)d\tau_1.\quad (25)$$

Viewed as a steady state problem we may wish, for example, to find that initial condition  $\phi_0$  (or better just the  $\psi$ ) which leads to a periodic motion of  $\psi$ . This is most practically done, of course, by solving the Dirac equation,

$$(i\nabla - m)\psi(1) = V(1)\psi(1),\quad (26)$$

deduced from (25) by operating on both sides by  $i\nabla_2 - m$ , thereby eliminating the  $\phi_0$ , and using (12). This illustrates the relation between the points of view.

For many problems the total potential  $\mathbf{A} + V$  may be split conveniently into a fixed one,  $V$ , and another,  $\mathbf{A}$ , considered as a perturbation. If  $K_+^{(V)}$  is defined as in

(16) with  $V$  for  $A$ , expressions such as (23) are valid and useful with  $K_+$  replaced by  $K_+^{(V)}$  and the functions  $f(1)$ ,  $g(2)$  replaced by solutions for all space and time of the Dirac Eq. (26) in the potential  $V$  (rather than free particle wave functions).

#### 4. PROBLEMS INVOLVING SEVERAL CHARGES

We wish next to consider the case that there are two (or more) distinct charges (in addition to pairs they may produce in virtual states). In a succeeding paper we discuss the interaction between such charges. Here we assume that they do not interact. In this case each particle behaves independently of the other. We can expect that if we have two particles  $a$  and  $b$ , the amplitude that particle  $a$  goes from  $x_1$  at  $t_1$ , to  $x_3$  at  $t_3$  while  $b$  goes from  $x_2$  at  $t_2$  to  $x_4$  at  $t_4$  is the product

$$K(3, 4; 1, 2) = K_{+a}(3, 1)K_{+b}(4, 2).$$

The symbols  $a$ ,  $b$  simply indicate that the matrices appearing in the  $K_+$  apply to the Dirac four component spinors corresponding to particle  $a$  or  $b$  respectively (the wave function now having 16 indices). In a potential  $K_{+a}$  and  $K_{+b}$  become  $K_{+a}^{(A)}$  and  $K_{+b}^{(A)}$  where  $K_{+a}^{(A)}$  is defined and calculated as for a single particle. They commute. Hereafter the  $a$ ,  $b$  can be omitted; the space time variable appearing in the kernels suffice to define on what they operate.

The particles are identical however and satisfy the exclusion principle. The principle requires only that one calculate  $K(3, 4; 1, 2) - K(4, 3; 1, 2)$  to get the net amplitude for arrival of charges at 3, 4. (It is normalized assuming that when an integral is performed over points 3 and 4, for example, since the electrons represented are identical, one divides by 2.) This expression is correct for positrons also (Fig. 4). For example the amplitude that an electron and a positron found initially at  $x_1$  and  $x_4$  (say  $t_1=t_4$ ) are later found at  $x_3$  and  $x_2$  (with  $t_2=t_3>t_1$ ) is given by the same expression

$$K_+^{(A)}(3, 1)K_+^{(A)}(4, 2) - K_+^{(A)}(4, 1)K_+^{(A)}(3, 2). \quad (27)$$

The first term represents the amplitude that the electron proceeds from 1 to 3 and the positron from 4 to 2 (Fig. 4(c)), while the second term represents the interfering amplitude that the pair at 1, 4 annihilate and what is found at 3, 2 is a pair newly created in the potential. The generalization to several particles is clear. There is an additional factor  $K_+^{(A)}$  for each particle, and anti-symmetric combinations are always taken.

No account need be taken of the exclusion principle in intermediate states. As an example consider again expression (14) for  $t_2>t_1$  and suppose  $t_4< t_3$  so that the situation represented (Fig. 2(c)) is that a pair is made at 4 with the electron proceeding to 2, and the positron to 3 where it annihilates the electron arriving from 1. It may be objected that if it happens that the electron created at 4 is in the same state as the one coming from 1, then the process cannot occur because of the exclusion principle and we should not have included it in our

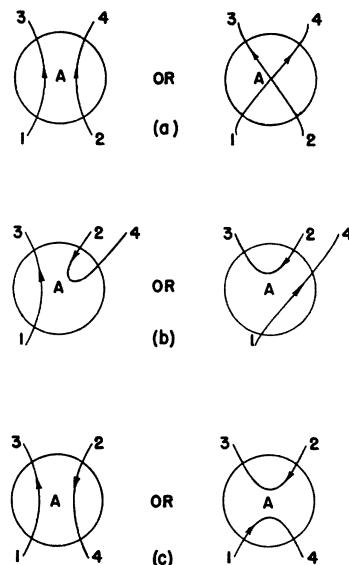


FIG. 4. Some problems involving two distinct charges (in addition to virtual pairs they may produce):  $P_v [K_+^{(A)}(3, 1)K_+^{(A)}(4, 2) - K_+^{(A)}(4, 1)K_+^{(A)}(3, 2)]^2$  is the probability that: (a) Electrons at 1 and 2 are scattered to 3, 4 (and no pairs are formed). (b) Starting with an electron at 1 a single pair is formed, positron at 2, electrons at 3, 4. (c) A pair at 1, 4 is found at 3, 2, etc. The exclusion principle requires that the amplitudes for processes involving exchange of two electrons be subtracted.

term (14). We shall see, however, that considering the exclusion principle also requires another change which reinstates the quantity.

For we are computing amplitudes relative to the amplitude that a vacuum at  $t_1$  will still be a vacuum at  $t_2$ . We are interested in the alteration in this amplitude due to the presence of an electron at 1. Now one process that can be visualized as occurring in the vacuum is the creation of a pair at 4 followed by a re-annihilation of the same pair at 3 (a process which we shall call a closed loop path). But if a real electron is present in a certain state 1, those pairs for which the electron was created in state 1 in the vacuum must now be excluded. We must therefore subtract from our relative amplitude the term corresponding to this process. But this just reinstates the quantity which it was argued should not have been included in (14), the necessary minus sign coming automatically from the definition of  $K_+$ . It is obviously simpler to disregard the exclusion principle completely in the intermediate states.

All the amplitudes are relative and their squares give the relative probabilities of the various phenomena. Absolute probabilities result if one multiplies each of the probabilities by  $P_v$ , the true probability that if one has no particles present initially there will be none finally. This quantity  $P_v$  can be calculated by normalizing the relative probabilities such that the sum of the probabilities of all mutually exclusive alternatives is unity. (For example if one starts with a vacuum one can calculate the relative probability that there remains a

vacuum (unity), or one pair is created, or two pairs, etc. The sum is  $P_v^{-1}$ .) Put in this form the theory is complete and there are no divergence problems. Real processes are completely independent of what goes on in the vacuum.

When we come, in the succeeding paper, to deal with interactions between charges, however, the situation is not so simple. There is the possibility that virtual electrons in the vacuum may interact electromagnetically with the real electrons. For that reason processes occurring in the vacuum are analyzed in the next section, in which an independent method of obtaining  $P_v$  is discussed.

### 5. VACUUM PROBLEMS

An alternative way of obtaining absolute amplitudes is to multiply all amplitudes by  $C_v$ , the vacuum to vacuum amplitude, that is, the absolute amplitude that there be no particles both initially and finally. We can assume  $C_v=1$  if no potential is present during the interval, and otherwise we compute it as follows. It differs from unity because, for example, a pair could be created which eventually annihilates itself again. Such a path would appear as a closed loop on a space-time diagram. The sum of the amplitudes resulting from all such single closed loops we call  $L$ . To a first approximation  $L$  is

$$L^{(1)} = -\frac{1}{2} \int \int Sp[K_+(2, 1)A(1) \\ \times K_+(1, 2)A(2)]d\tau_1 d\tau_2. \quad (28)$$

For a pair could be created say at 1, the electron and positron could both go on to 2 and there annihilate. The spur,  $Sp$ , is taken since one has to sum over all possible spins for the pair. The factor  $\frac{1}{2}$  arises from the fact that the same loop could be considered as starting at either potential, and the minus sign results since the interactors are each  $-iA$ . The next order term would be<sup>9</sup>

$$L^{(2)} = +(i/3) \int \int \int Sp[K_+(2, 1)A(1) \\ \times K_+(1, 3)A(3)K_+(3, 2)A(2)]d\tau_1 d\tau_2 d\tau_3,$$

etc. The sum of all such terms gives  $L^{(10)}$ .

<sup>9</sup> This term actually vanishes as can be seen as follows. In any spur the sign of all  $\gamma$  matrices may be reversed. Reversing the sign of  $\gamma$  in  $K_+(2, 1)$  changes it to the transpose of  $K_+(1, 2)$  so that the order of all factors and variables is reversed. Since the integral is taken over all  $\tau_1$ ,  $\tau_2$ , and  $\tau_3$  this has no effect and we are left with  $(-1)^3$  from changing the sign of  $A$ . Thus the spur equals its negative. Loops with an odd number of potential interactors give zero. Physically this is because for each loop the electron can go around one way or in the opposite direction and we must add these amplitudes. But reversing the motion of an electron makes it behave like a positive charge thus changing the sign of each potential interaction, so that the sum is zero if the number of interactions is odd. This theorem is due to W. H. Furry, Phys. Rev. 51, 125 (1937).

<sup>10</sup> A closed expression for  $L$  in terms of  $K_+^{(A)}$  is hard to obtain because of the factor  $(1/n)$  in the  $n$ th term. However, the perturbation in  $L$ ,  $\Delta L$  due to a small change in potential  $\Delta A$ , is easy to express. The  $(1/n)$  is canceled by the fact that  $\Delta A$  can appear

In addition to these single loops we have the possibility that two independent pairs may be created and each pair may annihilate itself again. That is, there may be formed in the vacuum two closed loops, and the contribution in amplitude from this alternative is just the product of the contribution from each of the loops considered singly. The total contribution from all such pairs of loops (it is still consistent to disregard the exclusion principle for these virtual states) is  $L^2/2$  for in  $L^2$  we count every pair of loops twice. The total vacuum-vacuum amplitude is then

$$C_v = 1 - L + L^2/2 - L^3/6 + \dots = \exp(-L), \quad (30)$$

the successive terms representing the amplitude from zero, one, two, etc., loops. The fact that the contribution to  $C_v$  of single loops is  $-L$  is a consequence of the Pauli principle. For example, consider a situation in which two pairs of particles are created. Then these pairs later destroy themselves so that we have two loops. The electrons could, at a given time, be interchanged forming a kind of figure eight which is a single loop. The fact that the interchange must change the sign of the contribution requires that the terms in  $C_v$  appear with alternate signs. (The exclusion principle is also responsible in a similar way for the fact that the amplitude for a pair creation is  $-K_+$  rather than  $+K_+$ .) Symmetrical statistics would lead to

$$C_v = 1 + L + L^2/2 = \exp(+L).$$

The quantity  $L$  has an infinite imaginary part (from  $L^{(1)}$ , higher orders are finite). We will discuss this in connection with vacuum polarization in the succeeding paper. This has no effect on the normalization constant for the probability that a vacuum remain vacuum is given by

$$P_v = |C_v|^2 = \exp(-2 \cdot \text{real part of } L),$$

from (30). This value agrees with the one calculated directly by renormalizing probabilities. The real part of  $L$  appears to be positive as a consequence of the Dirac equation and properties of  $K_+$  so that  $P_v$  is less than one. Bose statistics gives  $C_v = \exp(+L)$  and consequently a value of  $P_v$  greater than unity which appears meaningless if the quantities are interpreted as we have done here. Our choice of  $K_+$  apparently requires the exclusion principle.

Charges obeying the Klein-Gordon equation can be equally well treated by the methods which are discussed here for the Dirac electrons. How this is done is discussed in more detail in the succeeding paper. The real part of  $L$  comes out negative for this equation so that in this case Bose statistics appear to be required for consistency.<sup>8</sup>

in any of the  $n$  potentials. The result after summing over  $n$  by (13), (14) and using (16) is

$$\Delta L = -i \int Sp[(K_+^{(A)}(1, 1) - K_+(1, 1)) \Delta A(1)] d\tau_1. \quad (29)$$

The term  $K_+(1, 1)$  actually integrates to zero.

## 6. ENERGY-MOMENTUM REPRESENTATION

The practical evaluation of the matrix elements in some problems is often simplified by working with momentum and energy variables rather than space and time. This is because the function  $K_+(2, 1)$  is fairly complicated but we shall find that its Fourier transform is very simple, namely  $(i/4\pi^2)(\mathbf{p}-\mathbf{m})^{-1}$  that is

$$K_+(2, 1) = (i/4\pi^2) \int (\mathbf{p}-\mathbf{m})^{-1} \exp(-i\mathbf{p} \cdot \mathbf{x}_{21}) d^4 p, \quad (31)$$

where  $\mathbf{p} \cdot \mathbf{x}_{21} = \mathbf{p} \cdot \mathbf{x}_2 - \mathbf{p} \cdot \mathbf{x}_1 = p_\mu x_{2\mu} - p_\mu x_{1\mu}$ ,  $\mathbf{p} = p_\mu \gamma_\mu$ , and  $d^4 p$  means  $(2\pi)^{-2} dp_1 dp_2 dp_3 dp_4$ , the integral over all  $p$ . That this is true can be seen immediately from (12), for the representation of the operator  $i\nabla - m$  in energy ( $p_4$ ) and momentum ( $p_{1,2,3}$ ) space is  $\mathbf{p} - \mathbf{m}$  and the transform of  $\delta(2, 1)$  is a constant. The reciprocal matrix  $(\mathbf{p} - \mathbf{m})^{-1}$  can be interpreted as  $(\mathbf{p} + \mathbf{m})(\mathbf{p}^2 - m^2)^{-1}$  for  $\mathbf{p}^2 - m^2 = (\mathbf{p} - \mathbf{m})(\mathbf{p} + \mathbf{m})$  is a pure number not involving  $\gamma$  matrices. Hence if one wishes one can write

$$K_+(2, 1) = i(i\nabla_2 + m) I_+(2, 1),$$

where

$$I_+(2, 1) = (2\pi)^{-2} \int (\mathbf{p}^2 - m^2)^{-1} \exp(-i\mathbf{p} \cdot \mathbf{x}_{21}) d^4 p, \quad (32)$$

is not a matrix operator but a function satisfying

$$\square_2^2 I_+(2, 1) - m^2 I_+(2, 1) = \delta(2, 1), \quad (33)$$

where  $-\square_2^2 = (\nabla_2)^2 = (\partial/\partial x_{2\mu})(\partial/\partial x_{2\mu})$ .

The integrals (31) and (32) are not yet completely defined for there are poles in the integrand when  $\mathbf{p}^2 - m^2 = 0$ . We can define how these poles are to be evaluated by the rule that  $m$  is considered to have an infinitesimal negative imaginary part. That is  $m$ , is replaced by  $m - i\delta$  and the limit taken as  $\delta \rightarrow 0$  from above. This can be seen by imagining that we calculate  $K_+$  by integrating on  $p_4$  first. If we call  $E = +(\mathbf{m}^2 + \mathbf{p}_1^2 + \mathbf{p}_2^2 + \mathbf{p}_3^2)^{\frac{1}{2}}$  then the integrals involve  $p_4$  essentially as  $\int \exp(-i\mathbf{p}_4(t_2 - t_1)) d\mathbf{p}_4 (\mathbf{p}_4^2 - E^2)^{-1}$  which has poles at  $\mathbf{p}_4 = +E$  and  $\mathbf{p}_4 = -E$ . The replacement of  $m$  by  $m - i\delta$  means that  $E$  has a small negative imaginary part; the first pole is below, the second above the real axis. Now if  $t_2 - t_1 > 0$  the contour can be completed around the semicircle below the real axis thus giving a residue from the  $\mathbf{p}_4 = +E$  pole, or  $-(2E)^{-1} \exp(-iE(t_2 - t_1))$ . If  $t_2 - t_1 < 0$  the upper semicircle must be used, and  $\mathbf{p}_4 = -E$  at the pole, so that the function varies in each case as required by the other definition (17).

Other solutions of (12) result from other prescriptions. For example if  $p_4$  in the factor  $(\mathbf{p}^2 - m^2)^{-1}$  is considered to have a positive imaginary part  $K_+$  becomes replaced by  $K_0$ , the Dirac one-electron kernel, zero for  $t_2 < t_1$ . Explicitly the function is<sup>11</sup>  $(\mathbf{x}, t = \mathbf{x}_{21\mu})$

$$I_+(\mathbf{x}, t) = -(4\pi)^{-1} \delta(s^2) + (m/8\pi s) H_1^{(2)}(ms), \quad (34)$$

where  $s = +(\mathbf{t}^2 - \mathbf{x}^2)^{\frac{1}{2}}$  for  $\mathbf{t}^2 > \mathbf{x}^2$  and  $s = -i(x^2 - t^2)^{\frac{1}{2}}$  for

<sup>11</sup>  $I_+(\mathbf{x}, t)$  is  $(2i)^{-1}(D_1(\mathbf{x}, t) - iD(\mathbf{x}, t))$  where  $D_1$  and  $D$  are the functions defined by W. Pauli, Rev. Mod. Phys. 13, 203 (1941).

$\mathbf{t}^2 < \mathbf{x}^2$ ,  $H_1^{(2)}$  is the Hankel function and  $\delta(s^2)$  is the Dirac delta function of  $s^2$ . It behaves asymptotically as  $\exp(-ims)$ , decaying exponentially in space-like directions.<sup>12</sup>

By means of such transforms the matrix elements like (22), (23) are easily worked out. A free particle wave function for an electron of momentum  $\mathbf{p}_1$  is  $u_1 \exp(-i\mathbf{p}_1 \cdot \mathbf{x})$  where  $u_1$  is a constant spinor satisfying the Dirac equation  $\mathbf{p}_1 u_1 = mu_1$  so that  $\mathbf{p}_1^2 = m^2$ . The matrix element (22) for going from a state  $\mathbf{p}_1, u_1$  to a state of momentum  $\mathbf{p}_2$ , spinor  $u_2$ , is  $-4\pi^2 i (\bar{u}_2 \mathbf{a}(\mathbf{q}) u_1)$  where we have imagined  $\mathbf{A}$  expanded in a Fourier integral

$$A(1) = \int \mathbf{a}(\mathbf{q}) \exp(-i\mathbf{q} \cdot \mathbf{x}_1) d^4 q,$$

and we select the component of momentum  $\mathbf{q} = \mathbf{p}_2 - \mathbf{p}_1$ .

The second order term (23) is the matrix element between  $u_1$  and  $u_2$  of

$$-4\pi^2 i \int (\mathbf{a}(\mathbf{p}_2 - \mathbf{p}_1 - \mathbf{q})) (\mathbf{p}_1 + \mathbf{q} - \mathbf{m})^{-1} \mathbf{a}(\mathbf{q}) d^4 q, \quad (35)$$

since the electron of momentum  $\mathbf{p}_1$  may pick up  $\mathbf{q}$  from the potential  $\mathbf{a}(\mathbf{q})$ , propagate with momentum  $\mathbf{p}_1 + \mathbf{q}$  (factor  $(\mathbf{p}_1 + \mathbf{q} - \mathbf{m})^{-1}$ ) until it is scattered again by the potential,  $\mathbf{a}(\mathbf{p}_2 - \mathbf{p}_1 - \mathbf{q})$ , picking up the remaining momentum,  $\mathbf{p}_2 - \mathbf{p}_1 - \mathbf{q}$ , to bring the total to  $\mathbf{p}_2$ . Since all values of  $\mathbf{q}$  are possible, one integrates over  $\mathbf{q}$ .

These same matrices apply directly to positron problems, for if the time component of, say,  $\mathbf{p}_1$  is negative the state represents a positron of four-momentum  $-\mathbf{p}_1$ , and we are describing pair production if  $\mathbf{p}_2$  is an electron, i.e., has positive time component, etc.

The probability of an event whose matrix element is  $(\bar{u}_2 M u_1)$  is proportional to the absolute square. This may also be written  $(\bar{u}_1 \bar{M} u_2)(\bar{u}_2 M u_1)$ , where  $\bar{M}$  is  $M$  with the operators written in opposite order and explicit appearance of  $i$  changed to  $-i$  ( $\bar{M}$  is  $\beta$  times the complex conjugate transpose of  $\beta M$ ). For many problems we are not concerned about the spin of the final state. Then we can sum the probability over the two  $u_2$  corresponding to the two spin directions. This is not a complete set because  $\mathbf{p}_2$  has another eigenvalue,  $-m$ . To permit summing over all states we can insert the projection operator  $(2m)^{-1}(\mathbf{p}_2 + m)$  and so obtain  $(2m)^{-1}(\bar{u}_1 \bar{M}(\mathbf{p}_2 + m) M u_1)$  for the probability of transition from  $\mathbf{p}_1, u_1$ , to  $\mathbf{p}_2$  with arbitrary spin. If the incident state is unpolarized we can sum on its spins too, and obtain

$$(2m)^{-2} S_p [(\mathbf{p}_1 + m) \bar{M} (\mathbf{p}_2 + m) M] \quad (36)$$

for (twice) the probability that an electron of arbitrary spin with momentum  $\mathbf{p}_1$  will make transition to  $\mathbf{p}_2$ . The expressions are all valid for positrons when  $\mathbf{p}$ 's with

<sup>12</sup> If the  $-i\delta$  is kept with  $m$  here too the function  $I_+$  approaches zero for infinite positive and negative times. This may be useful in general analyses in avoiding complications from infinitely remote surfaces.

negative energies are inserted, and the situation interpreted in accordance with the timing relations discussed above. (We have used functions normalized to  $\langle \bar{u}u \rangle = 1$  instead of the conventional  $\langle \bar{u}\beta u \rangle = \langle u^*u \rangle = 1$ . On our scale  $\langle \bar{u}\beta u \rangle = \text{energy}/m$  so the probabilities must be corrected by the appropriate factors.)

The author has many people to thank for fruitful conversations about this subject, particularly H. A. Bethe and F. J. Dyson.

## APPENDIX

### a. Deduction from Second Quantization

In this section we shall show the equivalence of this theory with the hole theory of the positron.<sup>2</sup> According to the theory of second quantization of the electron field in a given potential,<sup>13</sup> the state of this field at any time is represented by a wave function  $\chi$  satisfying

$$i\partial\chi/\partial t = H\chi,$$

where  $H = \int \Psi^*(\mathbf{x}) (\alpha \cdot (-i\nabla - \mathbf{A}) + A_4 + m\beta) \Psi(\mathbf{x}) d^3x$  and  $\Psi(\mathbf{x})$  is an operator annihilating an electron at position  $\mathbf{x}$ , while  $\Psi^*(\mathbf{x})$  is the corresponding creation operator. We contemplate a situation in which at  $t=0$  we have present some electrons in states represented by ordinary spinor functions  $f_1(\mathbf{x}), f_2(\mathbf{x}), \dots$  assumed orthogonal, and some positrons. These are described as holes in the negative energy sea, the electrons which would normally fill the holes having wave functions  $p_1(\mathbf{x}), p_2(\mathbf{x}), \dots$ . We ask, at time  $T$  what is the amplitude that we find electrons in states  $g_1(\mathbf{x}), g_2(\mathbf{x}), \dots$  and holes at  $q_1(\mathbf{x}), q_2(\mathbf{x}), \dots$ . If the initial and final state vectors representing this situation are  $\chi_i$  and  $\chi_f$  respectively, we wish to calculate the matrix element

$$R = \left( \chi_f^* \exp\left(-i \int_0^T H dt\right) \chi_i \right) = \langle \chi_f^* S \chi_i \rangle. \quad (37)$$

We assume that the potential  $A$  differs from zero only for times between 0 and  $T$  so that a vacuum can be defined at these times. If  $\chi_0$  represents the vacuum state (that is, all negative energy states filled, all positive energies empty), the amplitude for having a vacuum at time  $T$ , if we had one at  $t=0$ , is

$$C_v = \langle \chi_0^* S \chi_0 \rangle, \quad (38)$$

writing  $S$  for  $\exp(-i \int_0^T H dt)$ . Our problem is to evaluate  $R$  and show that it is a simple factor times  $C_v$ , and that the factor involves the  $K_+^{(A)}$  functions in the way discussed in the previous sections.

To do this we first express  $\chi_i$  in terms of  $\chi_0$ . The operator

$$\Phi^* = \int \Psi^*(\mathbf{x}) \phi(\mathbf{x}) d^3x, \quad (39)$$

creates an electron with wave function  $\phi(\mathbf{x})$ . Likewise  $\Phi = \int \phi^*(\mathbf{x}) \times \Psi(\mathbf{x}) d^3x$  annihilates one with wave function  $\phi(\mathbf{x})$ . Hence state  $\chi_i$  is  $\chi_i = F_1^* F_2^* \dots P_1 P_2 \dots \chi_0$  while the final state is  $G_1^* G_2^* \dots Q_1 Q_2 \dots \chi_0$  where  $F_i, G_i, P_i, Q_i$  are operators defined like  $\Phi$ , in (39), but with  $f_i, g_i, p_i, q_i$  replacing  $\phi$ ; for the initial state would result from the vacuum if we created the electrons in  $f_1, f_2, \dots$  and annihilated those in  $p_1, p_2, \dots$ . Hence we must find

$$R = \langle \chi_0^* \dots Q_2^* Q_1^* \dots G_2 G_1 S F_1^* F_2^* \dots P_1 P_2 \dots \chi_0 \rangle. \quad (40)$$

To simplify this we shall have to use commutation relations between a  $\Phi^*$  operator and  $S$ . To this end consider  $\exp(-i \int_0^t H dt) \Phi^* \times \exp(+i \int_0^t H dt')$  and expand this quantity in terms of  $\Psi^*(\mathbf{x})$ , giving  $\int \Psi^*(\mathbf{x}) \phi(\mathbf{x}, t) d^3x$ , (which defines  $\phi(\mathbf{x}, t)$ ). Now multiply this equation by  $\exp(+i \int_0^t H dt') \dots \exp(-i \int_0^t H dt')$  and find

$$\int \Psi^*(\mathbf{x}) \phi(\mathbf{x}) d^3x = \int \Psi^*(\mathbf{x}, t) \phi(\mathbf{x}, t) d^3x, \quad (41)$$

where we have defined  $\Psi(\mathbf{x}, t)$  by  $\Psi(\mathbf{x}, t) = \exp(+i \int_0^t H dt') \Psi(\mathbf{x})$

<sup>13</sup> See, for example, G. Wentzel, *Einführung in die Quantentheorie der Wellenfelder* (Franz Deuticke, Leipzig, 1943), Chapter V.

$\times \exp(-i \int_0^t H dt')$ . As is well known  $\Psi(\mathbf{x}, t)$  satisfies the Dirac equation, (differentiate  $\Psi(\mathbf{x}, t)$  with respect to  $t$  and use commutation relations of  $H$  and  $\Psi$ )

$$i\partial\Psi(\mathbf{x}, t)/\partial t = (\alpha \cdot (-i\nabla - \mathbf{A}) + A_4 + m\beta)\Psi(\mathbf{x}, t). \quad (42)$$

Consequently  $\phi(\mathbf{x}, t)$  must also satisfy the Dirac equation (differentiate (41) with respect to  $t$ , use (42) and integrate by parts).

That is, if  $\phi(\mathbf{x}, T)$  is that solution of the Dirac equation at time  $T$  which is  $\phi(\mathbf{x})$  at  $t=0$ , and if we define  $\Phi^* = \int \Psi^*(\mathbf{x}) \phi(\mathbf{x}) d^3x$  and  $\Phi'^* = \int \Psi^*(\mathbf{x}) \phi(\mathbf{x}, T) d^3x$  then  $\Phi'^* = S \Phi^* S^{-1}$ , or

$$S \Phi^* = \Phi'^* S. \quad (43)$$

The principle on which the proof will be based can now be illustrated by a simple example. Suppose we have just one electron initially and finally and ask for

$$r = \langle \chi_0^* G S \chi_0 \rangle. \quad (44)$$

We might try putting  $F^*$  through the operator  $S$  using (43),  $S F^* = F'^* S$ , where  $f'$  in  $F'^* = \int \Psi^*(\mathbf{x}) f'(\mathbf{x}) d^3x$  is the wave function at  $T$  arising from  $f(\mathbf{x})$  at 0. Then

$$r = \langle \chi_0^* G F'^* S \chi_0 \rangle = \int g^*(\mathbf{x}) f'(\mathbf{x}) d^3x \cdot C_v - \langle \chi_0^* F'^* G S \chi_0 \rangle, \quad (45)$$

where the second expression has been obtained by use of the definition (38) of  $C_v$  and the general commutation relation

$$G F^* + F^* G = \int g^*(\mathbf{x}) f(\mathbf{x}) d^3x,$$

which is a consequence of the properties of  $\Psi(\mathbf{x})$  (the others are  $FG = -GF$  and  $F^*G^* = -G^*F^*$ ). Now  $\chi_0^* F'^*$  in the last term in (45) is the complex conjugate of  $F' \chi_0$ . Thus if  $f'$  contained only positive energy components,  $F' \chi_0$  would vanish and we would have reduced  $r$  to a factor times  $C_v$ . But  $F'$ , as worked out here, does contain negative energy components created in the potential  $A$  and the method must be slightly modified.

Before putting  $F^*$  through the operator we shall add to it another operator  $F''^*$  arising from a function  $f''(\mathbf{x})$  containing *only negative* energy components and so chosen that the resulting  $f'$  has *only positive* ones. That is we want

$$S(F_{\text{pos}}^* + F_{\text{neg}}''^*) = F_{\text{pos}}'^* S, \quad (46)$$

where the “pos” and “neg” serve as reminders of the sign of the energy components contained in the operators. This we can now use in the form

$$S F_{\text{pos}}^* = F_{\text{pos}}'^* S - S F_{\text{neg}}''^*. \quad (47)$$

In our one electron problem this substitution replaces  $r$  by two terms

$$r = \langle \chi_0^* G F_{\text{pos}}'^* S \chi_0 \rangle - \langle \chi_0^* G S F_{\text{neg}}''^* \chi_0 \rangle.$$

The first of these reduces to

$$r = \int g^*(\mathbf{x}) f_{\text{pos}}'(\mathbf{x}) d^3x \cdot C_v,$$

as above, for  $F_{\text{pos}}' \chi_0$  is now zero, while the second is zero since the creation operator  $F_{\text{neg}}''^*$  gives zero when acting on the vacuum state as all negative energies are full. This is the central idea of the demonstration.

The problem presented by (46) is this: Given a function  $f_{\text{pos}}(\mathbf{x})$  at time 0, to find the amount,  $f_{\text{neg}}''$ , of negative energy component which must be added in order that the solution of Dirac's equation at time  $T$  will have only positive energy components,  $f_{\text{pos}}'$ . This is a boundary value problem for which the kernel  $K_+^{(A)}$  is designed. We know the positive energy components initially,  $f_{\text{pos}}$ , and the negative ones finally (zero). The positive ones finally are therefore (using (19))

$$f_{\text{pos}}'(\mathbf{x}_2) = \int K_+^{(A)}(2, 1) \beta f_{\text{pos}}(\mathbf{x}_1) d^3x_1, \quad (48)$$

where  $t_2 = T$ ,  $t_1 = 0$ . Similarly, the negative ones initially are

$$f_{\text{neg}}''(\mathbf{x}_2) = \int K_+^{(A)}(2, 1) \beta f_{\text{pos}}(\mathbf{x}_1) d^3x_1 - f_{\text{pos}}(\mathbf{x}_2), \quad (49)$$

where  $t_2$  approaches zero from above, and  $t_1 = 0$ . The  $f_{\text{pos}}(\mathbf{x}_2)$  is

subtracted to keep in  $f_{\text{neg}}''(\mathbf{x}_2)$  only those waves which return from the potential and not those arriving directly at  $t_2$  from the  $K_+(2, 1)$  part of  $K_+^{(A)}(2, 1)$ , as  $t_2 \rightarrow 0$ . We could also have written

$$f_{\text{neg}}''(\mathbf{x}_2) = \int [K_+^{(A)}(2, 1) - K_+(2, 1)] \beta f_{\text{pos}}(\mathbf{x}_1) d^3 \mathbf{x}_1. \quad (50)$$

Therefore the one-electron problem,  $r = \mathcal{F} g^*(\mathbf{x}) f_{\text{pos}}'(\mathbf{x}) d^3 \mathbf{x} \cdot C_v$ , gives by (48)

$$r = C_v \int g^*(\mathbf{x}_2) K_+^{(A)}(2, 1) \beta f(\mathbf{x}_1) d^3 \mathbf{x}_1 d^3 \mathbf{x}_2,$$

as expected in accordance with the reasoning of the previous sections (i.e., (20) with  $K_+^{(A)}$  replacing  $K_+$ ).

The proof is readily extended to the more general expression  $R$ , (40), which can be analyzed by induction. First one replaces  $F_1^*$  by a relation such as (47) obtaining two terms

$$R = (\chi_0^* \cdots Q_1^* Q_2^* \cdots G_2 G_1 F_{1\text{pos}}'^* S F_2^* \cdots P_1 P_2 \cdots \chi_0) \\ - (\chi_0^* \cdots Q_2^* Q_1^* \cdots G_2 G_1 S F_{1\text{neg}}'^* F_2^* \cdots P_1 P_2 \cdots \chi_0).$$

In the first term the order of  $F_{1\text{pos}}'^*$  and  $G_1$  is then interchanged, producing an additional term  $\mathcal{F} g_1^*(\mathbf{x}) f_{1\text{pos}}'(\mathbf{x}) d^3 \mathbf{x}$  times an expression with one less electron in initial and final state. Next it is exchanged with  $G_2$  producing an addition  $- \mathcal{F} g_2^*(\mathbf{x}) f_{1\text{pos}}'(\mathbf{x}) d^3 \mathbf{x}$  times a similar term, etc. Finally on reaching the  $Q_i^*$  with which it anticommutes it can be simply moved over to juxtaposition with  $\chi_0^*$  where it gives zero. The second term is similarly handled by moving  $F_{1\text{neg}}'^*$  through anti commuting  $F_2^*$ , etc., until it reaches  $P_1$ . Then it is exchanged with  $P_1$  to produce an additional simpler term with a factor  $\mp \mathcal{F} p_1^*(\mathbf{x}) f_{1\text{neg}}''(\mathbf{x}) d^3 \mathbf{x}$  or  $\mp \mathcal{F} p_1^*(\mathbf{x}_2) K_+^{(A)}(2, 1) \beta f_1(\mathbf{x}_1) d^3 \mathbf{x}_1 d^3 \mathbf{x}_2$  from (49), with  $t_2 = t_1 = 0$  (the extra  $f_1(\mathbf{x}_2)$  in (49) gives zero as it is orthogonal to  $p_1(\mathbf{x}_2)$ ). This describes in the expected manner the annihilation of the pair, electron  $f_1$ , positron  $p_1$ . The  $F_{\text{neg}}'^*$  is moved in this way successively through the  $P$ 's until it gives zero when acting on  $\chi_0$ . Thus  $R$  is reduced, with the expected factors (and with alternating signs as required by the exclusion principle), to simpler terms containing two less operators which may in turn be further reduced by using  $F_2^*$  in a similar manner, etc. After all the  $F^*$  are used the  $Q^*$ 's can be reduced in a similar manner. They are moved through the  $S$  in the opposite direction in such a manner as to produce a purely negative energy operator at time 0, using relations analogous to (46) to (49). After all this is done we are left simply with the expected factor times  $C_v$  (assuming the net charge is the same in initial and final state.)

In this way we have written the solution to the general problem of the motion of electrons in given potentials. The factor  $C_v$  is obtained by normalization. However for photon fields it is desirable to have an explicit form for  $C_v$  in terms of the potentials. This is given by (30) and (29) and it is readily demonstrated that this also is correct according to second quantization.

### b. Analysis of the Vacuum Problem

We shall calculate  $C_v$  from second quantization by induction considering a series of problems each containing a potential distribution more nearly like the one we wish. Suppose we know  $C_v$  for a problem like the one we want and having the same potentials for time  $t$  between some  $t_0$  and  $T$ , but having potential zero for times from 0 to  $t_0$ . Call this  $C_v(t_0)$ , the corresponding Hamiltonian  $H_{t_0}$  and the sum of contributions for all single loops,  $L(t_0)$ . Then for  $t_0 = T$  we have zero potential at all times, no pairs can be produced,  $L(T) = 0$  and  $C_v(T) = 1$ . For  $t_0 = 0$  we have the complete problem, so that  $C_v(0)$  is what is defined as  $C_v$  in (38). Generally we have,

$$C_v(t_0) = \left( \chi_0^* \exp \left( -i \int_{t_0}^T H_{t_0} dt \right) \chi_0 \right) \\ = \left( \chi_0^* \exp \left( -i \int_{t_0}^T H_{t_0} dt \right) \chi_0 \right),$$

since  $H_{t_0}$  is identical to the constant vacuum Hamiltonian  $H_T$  for  $t < t_0$  and  $\chi_0$  is an eigenfunction of  $H_T$  with an eigenvalue (energy of vacuum) which we can take as zero.

The value of  $C_v(t_0 - \Delta t_0)$  arises from the Hamiltonian  $H_{t_0 - \Delta t_0}$  which differs from  $H_{t_0}$  just by having an extra potential during the short interval  $\Delta t_0$ . Hence, to first order in  $\Delta t_0$ , we have

$$C_v(t_0 - \Delta t_0) = \left( \chi_0^* \exp \left( -i \int_{t_0 - \Delta t_0}^T H_{t_0 - \Delta t_0} dt \right) \chi_0 \right) \\ = \left( \chi_0^* \exp \left( -i \int_{t_0}^T H_{t_0} dt \right) \left[ 1 - i \Delta t_0 \int \Psi^*(\mathbf{x}) \right. \right. \\ \times \left. \left. (-\alpha \cdot \mathbf{A}(\mathbf{x}, t_0) + A_4(\mathbf{x}, t_0)) \Psi(\mathbf{x}) d^3 \mathbf{x} \right] \chi_0 \right);$$

we therefore obtain for the derivative of  $C_v$  the expression

$$-dC_v(t_0)/dt_0 = -i \left( \chi_0^* \exp \left( -i \int_{t_0}^T H_{t_0} dt \right) \right. \\ \times \left. \int \Psi^*(\mathbf{x}) \beta \mathbf{A}(\mathbf{x}, t_0) \Psi(\mathbf{x}) d^3 \mathbf{x} \chi_0 \right), \quad (51)$$

which will be reduced to a simple factor times  $C_v(t_0)$  by methods analogous to those used in reducing  $R$ . The operator  $\Psi$  can be imagined to be split into two pieces  $\Psi_{\text{pos}}$  and  $\Psi_{\text{neg}}$  operating on positive and negative energy states respectively. The  $\Psi_{\text{pos}}$  on  $\chi_0$  gives zero so we are left with two terms in the current density,  $\Psi_{\text{pos}}^* \beta \mathbf{A} \Psi_{\text{neg}}$  and  $\Psi_{\text{neg}}^* \beta \mathbf{A} \Psi_{\text{neg}}$ . The latter  $\Psi_{\text{neg}}^* \beta \mathbf{A} \Psi_{\text{neg}}$  is just the expectation value of  $\beta \mathbf{A}$  taken over all negative energy states (minus  $\Psi_{\text{neg}} \beta \mathbf{A} \Psi_{\text{neg}}^*$  which gives zero acting on  $\chi_0$ ). This is the effect of the vacuum expectation current of the electrons in the sea which we should have subtracted from our original Hamiltonian in the customary way.

The remaining term  $\Psi_{\text{pos}}^* \beta \mathbf{A} \Psi_{\text{neg}}$ , or its equivalent  $\Psi_{\text{pos}}^* \beta \mathbf{A} \Psi$  can be considered as  $\Psi^*(\mathbf{x}) f_{\text{pos}}(\mathbf{x})$  where  $f_{\text{pos}}(\mathbf{x})$  is written for the positive energy component of the operator  $\beta \mathbf{A} \Psi(\mathbf{x})$ . Now this operator,  $\Psi^*(\mathbf{x}) f_{\text{pos}}(\mathbf{x})$ , or more precisely just the  $\Psi^*(\mathbf{x})$  part of it, can be pushed through the  $\exp(-i \int_{t_0}^T H dt)$  in a manner exactly analogous to (47) when  $f$  is a function. (An alternative derivation results from the consideration that the operator  $\Psi(\mathbf{x}, t)$  which satisfies the Dirac equation also satisfies the linear integral equations which are equivalent to it.) That is, (51) can be written by (48), (50),

$$-dC_v(t_0)/dt_0 = -i \left( \chi_0^* \int \int \Psi^*(\mathbf{x}_2) K_+^{(A)}(2, 1) \right. \\ \times \exp \left( -i \int_{t_0}^T H dt \right) \mathbf{A}(1) \Psi(\mathbf{x}_1) d^3 \mathbf{x}_1 d^3 \mathbf{x}_2 \chi_0 \Big) \\ + i \left( \chi_0^* \exp \left( -i \int_{t_0}^T H dt \right) \int \int \Psi^*(\mathbf{x}_2) [K_+^{(A)}(2, 1) \right. \\ \left. - K_+(2, 1)] \mathbf{A}(1) \Psi(\mathbf{x}_1) d^3 \mathbf{x}_1 d^3 \mathbf{x}_2 \chi_0 \right),$$

where in the first term  $t_2 = T$ , and in the second  $t_2 \rightarrow t_0 = t_1$ . The  $(A)$  in  $K_+^{(A)}$  refers to that part of the potential  $A$  after  $t_0$ . The first term vanishes for it involves (from the  $K_+^{(A)}(2, 1)$ ) only positive energy components of  $\Psi^*$ , which give zero operating into  $\chi_0^*$ . In the second term only negative components of  $\Psi^*(\mathbf{x}_2)$  appear. If, then  $\Psi^*(\mathbf{x}_2)$  is interchanged in order with  $\Psi(\mathbf{x}_1)$  it will give zero operating on  $\chi_0$ , and only the term,

$$-dC_v(t_0)/dt_0 = +i \int Sp[(K_+^{(A)}(1, 1) \\ - K_+(1, 1)) \mathbf{A}(1)] d^3 \mathbf{x}_1 \cdot C_v(t_0), \quad (52)$$

will remain, from the usual commutation relation of  $\Psi^*$  and  $\Psi$ .

The factor of  $C_v(t_0)$  in (52) times  $-\Delta t_0$  is, according to (29) (reference 10), just  $L(t_0 - \Delta t_0) - L(t_0)$  since this difference arises from the extra potential  $\Delta \mathbf{A} = \mathbf{A}$  during the short time interval  $\Delta t_0$ . Hence  $-dC_v(t_0)/dt_0 = + (dL(t_0)/dt_0) C_v(t_0)$  so that integration from  $t_0 = T$  to  $t_0 = 0$  establishes (30).

Starting from the theory of the electromagnetic field in second quantization, a deduction of the equations for quantum electrodynamics which appear in the succeeding paper may be worked out using very similar principles. The Pauli-Weisskopf theory of the Klein-Gordon equation can apparently be analyzed in essentially the same way as that used here for Dirac electrons.



## Mathematical Formulation of the Quantum Theory of Electromagnetic Interaction

R. P. FEYNMAN\*

*Department of Physics, Cornell University, Ithaca, New York*

(Received June 8, 1950)

The validity of the rules given in previous papers for the solution of problems in quantum electrodynamics is established. Starting with Fermi's formulation of the field as a set of harmonic oscillators, the effect of the oscillators is integrated out in the Lagrangian form of quantum mechanics. There results an expression for the effect of all virtual photons valid to all orders in  $e^2/\hbar c$ . It is shown that evaluation of this expression as a power series in  $e^2/\hbar c$  gives just the terms expected by the aforementioned rules.

In addition, a relation is established between the amplitude for a given process in an arbitrary unquantized potential and in a quantum electrodynamical field. This relation permits a simple general statement of the laws of quantum electrodynamics.

A description, in Lagrangian quantum-mechanical form, of particles satisfying the Klein-Gordon equation is given in an Appendix. It involves the use of an extra parameter analogous to proper time to describe the trajectory of the particle in four dimensions.

A second Appendix discusses, in the special case of photons, the problem of finding what real processes are implied by the formula for virtual processes.

Problems of the divergences of electrodynamics are not discussed.

### 1. INTRODUCTION

**I**N two previous papers<sup>1</sup> rules were given for the calculation of the matrix element for any process in electrodynamics, to each order in  $e^2/\hbar c$ . No complete proof of the equivalence of these rules to the conventional electrodynamics was given in these papers. Secondly, no closed expression was given valid to all orders in  $e^2/\hbar c$ . In this paper these formal omissions will be remedied.<sup>2</sup>

In paper **II** it was pointed out that for many problems in electrodynamics the Hamiltonian method is not advantageous, and might be replaced by the over-all space-time point of view of a direct particle interaction. It was also mentioned that the Lagrangian form of quantum mechanics<sup>3</sup> was useful in this connection. The rules given in paper **II** were, in fact, first deduced in this form of quantum mechanics. We shall give this derivation here.

The advantage of a Lagrangian form of quantum mechanics is that in a system with interacting parts it permits a separation of the problem such that the motion of any part can be analyzed or solved first, and the results of this solution may then be used in the solution of the motion of the other parts. This separation is especially useful in quantum electrodynamics which represents the interaction of matter with the electromagnetic field. The electromagnetic field is an especially simple system and its behavior can be analyzed completely. What we shall show is that the

net effect of the field is a delayed interaction of the particles. It is possible to do this easily only if it is not necessary at the same time to analyze completely the motion of the particles. The only advantage in our problems of the form of quantum mechanics in **C** is to permit one to separate these aspects of the problem. There are a number of disadvantages, however, such as a lack of familiarity, the apparent (but not real) necessity for dealing with matter in non-relativistic approximation, and at times a cumbersome mathematical notation and method, as well as the fact that a great deal of useful information that is known about operators cannot be directly applied.

It is also possible to separate the field and particle aspects of a problem in a manner which uses operators and Hamiltonians in a way that is much more familiar. One abandons the notation that the order of action of operators depends on their written position on the paper and substitutes some other convention (such that the order of operators is that of the time to which they refer). The increase in manipulative facility which accompanies this change in notation makes it easier to represent and to analyze the formal problems in electrodynamics. The method requires some discussion, however, and will be described in a succeeding paper. In this paper we shall give the derivations of the formulas of **II** by means of the form of quantum mechanics given in **C**.

The problem of interaction of matter and field will be analyzed by first solving for the behavior of the field in terms of the coordinates of the matter, and finally discussing the behavior of the matter (by matter is actually meant the electrons and positrons). That is to say, we shall first eliminate the field variables from the equations of motion of the electrons and then discuss the behavior of the electrons. In this way all of the rules given in the paper **II** will be derived.

Actually, the straightforward elimination of the field

\* Now at the California Institute of Technology, Pasadena, California.

<sup>1</sup> R. P. Feynman, Phys. Rev. **76**, 749 (1949), hereafter called **I**, and Phys. Rev. **76**, 769 (1949), hereafter called **II**.

<sup>2</sup> See in this connection also the papers of S. Tomonaga, Phys. Rev. **74**, 224 (1948); S. Kanesawa and S. Tomonaga, Prog. Theoret. Phys. **3**, 101 (1948); J. Schwinger, Phys. Rev. **76**, 790 (1949); F. Dyson, Phys. Rev. **75**, 1736 (1949); W. Pauli and F. Villars, Rev. Mod. Phys. **21**, 434 (1949). The papers cited give references to previous work.

<sup>3</sup> R. P. Feynman, Rev. Mod. Phys. **20**, 367 (1948), hereafter called **C**.

variables will lead at first to an expression for the behavior of an arbitrary number of Dirac electrons. Since the number of electrons might be infinite, this can be used directly to find the behavior of the electrons according to hole theory by imagining that nearly all the negative energy states are occupied by electrons. But, at least in the case of motion in a fixed potential, it has been shown that this hole theory picture is equivalent to one in which a positron is represented as an electron whose space-time trajectory has had its time direction reversed. To show that this same picture may be used in quantum electrodynamics when the potentials are not fixed, a special argument is made based on a study of the relationship of quantum electrodynamics to motion in a fixed potential. Finally, it is pointed out that this relationship is quite general and might be used for a general statement of the laws of quantum electrodynamics.

Charges obeying the Klein-Gordon equation can be analyzed by a special formalism given in Appendix A. A fifth parameter is used to specify the four-dimensional trajectory so that the Lagrangian form of quantum mechanics can be used. Appendix B discusses in more detail the relation of real and virtual photon emission. An equation for the propagation of a self-interacting electron is given in Appendix C.

In the demonstration which follows we shall restrict ourselves temporarily to cases in which the particle's motion is non-relativistic, but the transition of the final formulas to the relativistic case is direct, and the proof could have been kept relativistic throughout.

The transverse part of the electromagnetic field will be represented as an assemblage of independent harmonic oscillators each interacting with the particles, as suggested by Fermi.<sup>4</sup> We use the notation of Heitler.<sup>5</sup>

## 2. QUANTUM ELECTRODYNAMICS IN LAGRANGIAN FORM

The Hamiltonian for a set of non-relativistic particles interacting with radiation is, classically,  $H = H_p + H_I + H_c + H_{tr}$ , where  $H_p + H_I = \sum_n \frac{1}{2} m_n^{-1} (\mathbf{p}_n - e_n \mathbf{A}^{tr}(\mathbf{x}_n))^2$  is the Hamiltonian of the particles of mass  $m_n$ , charge  $e_n$ , coordinate  $\mathbf{x}_n$  and momentum  $\mathbf{p}_n$  and their interaction with the transverse part of the electromagnetic field. This field can be expanded into plane waves

$$\mathbf{A}^{tr}(\mathbf{x}) = (8\pi)^{\frac{1}{2}} \sum_{\mathbf{K}} [\mathbf{e}_1(q_{\mathbf{K}}^{(1)} \cos(\mathbf{K} \cdot \mathbf{x}) + q_{\mathbf{K}}^{(3)} \sin(\mathbf{K} \cdot \mathbf{x})) + \mathbf{e}_2(q_{\mathbf{K}}^{(2)} \cos(\mathbf{K} \cdot \mathbf{x}) + q_{\mathbf{K}}^{(4)} \sin(\mathbf{K} \cdot \mathbf{x}))] \quad (1)$$

where  $\mathbf{e}_1$  and  $\mathbf{e}_2$  are two orthogonal polarization vectors at right angles to the propagation vector  $\mathbf{K}$ , magnitude  $k$ . The sum over  $\mathbf{K}$  means, if normalized to unit volume,  $\frac{1}{2} \int d^3 K / 8\pi^3$ , and each  $q_{\mathbf{K}}^{(r)}$  can be considered as the coordinate of a harmonic oscillator. (The factor  $\frac{1}{2}$  arises for the mode corresponding to  $\mathbf{K}$  and to  $-\mathbf{K}$  is the

same.) The Hamiltonian of the transverse field represented as oscillators is

$$H_{tr} = \frac{1}{2} \sum_{\mathbf{K}} \sum_{r=1}^4 ((p_{\mathbf{K}}^{(r)})^2 + k^2 (q_{\mathbf{K}}^{(r)})^2)$$

where  $p_{\mathbf{K}}^{(r)}$  is the momentum conjugate to  $q_{\mathbf{K}}^{(r)}$ . The longitudinal part of the field has been replaced by the Coulomb interaction,<sup>6</sup>

$$H_c = \frac{1}{2} \sum_n \sum_m e_n e_m / r_{nm}$$

where  $r_{nm}^2 = (\mathbf{x}_n - \mathbf{x}_m)^2$ . As is well known,<sup>4</sup> when this Hamiltonian is quantized one arrives at the usual theory of quantum electrodynamics. To express these laws of quantum electrodynamics one can equally well use the Lagrangian form of quantum mechanics to describe this set of oscillators and particles. The classical Lagrangian equivalent to this Hamiltonian is  $L = L_p + L_I + L_c + L_{tr}$  where

$$L_p = \frac{1}{2} \sum_n m_n \dot{\mathbf{x}}_n^2 \quad (2a)$$

$$L_I = \sum_n e_n \dot{\mathbf{x}}_n \cdot \mathbf{A}^{tr}(\mathbf{x}_n) \quad (2b)$$

$$L_{tr} = \frac{1}{2} \sum_{\mathbf{K}} \sum_r ((\dot{q}_{\mathbf{K}}^{(r)})^2 - k^2 (q_{\mathbf{K}}^{(r)})^2) \quad (2c)$$

$$L_c = -\frac{1}{2} \sum_n \sum_m e_n e_m / r_{nm}. \quad (2d)$$

When this Lagrangian is used in the Lagrangian forms of quantum mechanics of **C**, what it leads to is, of course, mathematically equivalent to the result of using the Hamiltonian  $H$  in the ordinary way, and is therefore equivalent to the more usual forms of quantum electrodynamics (at least for non-relativistic particles). We may, therefore, proceed by using this Lagrangian form of quantum electrodynamics, with the assurance that the results obtained must agree with those obtained from the more usual Hamiltonian form.

The Lagrangian enters through the statement that the functional which carries the system from one state to another is  $\exp(iS)$  where

$$S = \int L dt = S_p + S_I + S_c + S_{tr}. \quad (3)$$

The time integrals must be written as Riemann sums with some care; for example,

$$S_I = \sum_n \int e_n \dot{\mathbf{x}}_n(t) \cdot \mathbf{A}^{tr}(\mathbf{x}_n(t)) dt \quad (4)$$

becomes according to **C**, Eq. (19)

$$S_I = \sum_n \sum_i \frac{1}{2} e_n (\mathbf{x}_{n,i+1} - \mathbf{x}_{n,i}) \cdot (\mathbf{A}^{tr}(\mathbf{x}_{n,i+1}) + \mathbf{A}^{tr}(\mathbf{x}_{n,i})) \quad (5)$$

so that the velocity  $\dot{\mathbf{x}}_{n,i}$  which multiplies  $\mathbf{A}^{tr}(\mathbf{x}_{n,i})$  is

$$\dot{\mathbf{x}}_{n,i} = \frac{1}{2} \epsilon^{-1} (\mathbf{x}_{n,i+1} - \mathbf{x}_{n,i}) + \frac{1}{2} \epsilon^{-1} (\mathbf{x}_{n,i} - \mathbf{x}_{n,i-1}). \quad (6)$$

<sup>6</sup> The term in the sum for  $n=m$  is obviously infinite but must be included for relativistic invariance. Our problem here is to re-express the usual (and divergent) form of electrodynamics in the form given in **II**. Modifications for dealing with the divergences are discussed in **II** and we shall not discuss them further here.

<sup>4</sup> E. Fermi, Rev. Mod. Phys. 4, 87 (1932).  
<sup>5</sup> W. Heitler, *The Quantum Theory of Radiation*, second edition (Oxford University Press, London, 1944).

In the Lagrangian form it is possible to eliminate the transverse oscillators as is discussed in **C**, Section 13. One must specify, however, the initial and final state of all oscillators. We shall first choose the special, simple case that all oscillators are in their ground states initially and finally, so that all photons are virtual. Later we do the more general case in which real quanta are present initially or finally. We ask, then, for the amplitude for finding no quanta present and the particles in state  $\chi_{t''}$  at time  $t''$ , if at time  $t'$  the particles were in state  $\psi_{t'}$  and no quanta were present.

The method of eliminating field oscillators is described in Section 13 of **C**. We shall simply carry out the elimination here using the notation and equations of **C**. To do this, for simplicity, we first consider in the next section the case of a particle or a system of particles interacting with a single oscillator, rather than the entire assemblage of the electromagnetic field.

### 3. FORCED HARMONIC OSCILLATOR

We consider a harmonic oscillator, coordinate  $q$ , Lagrangian  $L = \frac{1}{2}(\dot{q}^2 - \omega^2 q^2)$  interacting with a particle or system of particles, action  $S_p$ , through a term in the Lagrangian  $q(t)\gamma(t)$  where  $\gamma(t)$  is a function of the coordinates (symbolized as  $x$ ) of the particle. The precise form of  $\gamma(t)$  for each oscillator of the electromagnetic field is given in the next section. We ask for the amplitude that at some time  $t''$  the particles are in state  $\chi_{t''}$  and the oscillator is in, say, an eigenstate  $m$  of energy  $\omega(m + \frac{1}{2})$  (units are chosen such that  $\hbar = c = 1$ ) when it is given that at a previous time  $t'$  the particles were in state  $\psi_{t'}$  and the oscillator in  $n$ . The amplitude for this is the transition amplitude [see **C**, Eq. (61)]

$$\begin{aligned} & \langle \chi_{t''} | \psi_{t'} \rangle_{S_p + S_0 + S_I} \\ &= \int \int \chi_{t''*}(x_{t''}) \varphi_m^*(q_{t''}) \exp(i(S_p + S_0 + S_I) \\ & \quad \varphi_n(q_{t'}) \psi_{t'}(x_{t'}) dx_{t'} dq_{t'} dq_{t''} \mathcal{D}x(t) \mathcal{D}q(t) \end{aligned} \quad (7)$$

where  $x$  represents the variables describing the particle,  $S_p$  is the action calculated classically for the particles for a given path going from coordinate  $x_{t'}$  at  $t'$  to  $x_{t''}$  at  $t''$ ,  $S_0$  is the action  $\int \frac{1}{2}(\dot{q}^2 - \omega^2 q^2) dt$  for any path of the oscillator going from  $q_{t'}$  at  $t'$  to  $q_{t''}$  at  $t''$ , while

$$S_I = \int q(t) \gamma(t) dt, \quad (8)$$

the action of interaction, is a functional of both  $q(t)$  and  $x(t)$ , the paths of oscillator and particles. The symbols  $\mathcal{D}x(t)$  and  $\mathcal{D}q(t)$  represent a summation over all possible paths of particles and oscillator which go between the given end points in the sense defined in **C**, Eq. (9). (That is, assuming time to proceed in infinitesimal steps,  $\epsilon$ , an integral over all values of the coordinates  $x$  and  $q$  corresponding to each instant in time, suitably normalized.)

The problem may be broken in two. The result can be written as an integral over all paths of the particles only, of  $(\exp i S_p) \cdot G_{mn}$ :

$$\langle \chi_{t''} | \psi_{t'} \rangle_{S_p + S_0 + S_I} = \langle \chi_{t''} | G_{mn} | \psi_{t'} \rangle_{S_p} \quad (9)$$

where  $G_{mn}$  is a functional of the path of the particles alone (since it depends on  $\gamma(t)$ ) given by

$$\begin{aligned} G_{mn} &= \left\langle \varphi_m \left| \exp i \int q(t) \gamma(t) dt \right| \varphi_n \right\rangle_{S_0} \\ &= \int \varphi_m^*(q_{t''}) \exp(i(S_0 + S_I)) \varphi_n(q_{t'}) dq_{t'} dq_{t''} \mathcal{D}q(t) \\ &= \int \varphi_m^*(q_j) \exp i \sum_{i=0}^{j-1} [\frac{1}{2} \epsilon^{-2} (q_{i+1} - q_i)^2 - \frac{1}{2} \omega^2 q_i^2 + q_i \gamma_i] \\ & \quad \cdot \varphi_n(q_0) dq_0 a^{-1} dq_1 a^{-1} dq_2 \cdots a^{-1} dq_j \end{aligned} \quad (10)$$

where we have written the  $\mathcal{D}q(t)$  out explicitly (and have set  $a = (2\pi i \epsilon)^{\frac{1}{2}}$ ,  $t'' - t' = j\epsilon$ ,  $q_{t'} = q_0$ ,  $q_{t''} = q_j$ ). The last form can be written as

$$G_{mn} = \int \varphi_m^*(q_j) k(q_j, t''; q_0, t') \varphi_n(q_0) dq_0 dq_j \quad (11)$$

where  $k(q_j, t''; q_0, t')$  is the kernel [as in **I**, Eq. (2)] for a forced harmonic oscillator giving the amplitude for arrival at  $q_j$  at time  $t''$  if at time  $t'$  it was known to be at  $q_0$ . According to **C** it is given by

$$k(q_j, t''; q_0, t') = (2\pi i \omega^{-1} \sin \omega(t'' - t'))^{-\frac{1}{2}} \exp i Q(q_j, t''; q_0, t') \quad (12)$$

where  $Q(q_j, t''; q_0, t')$  is the action calculated along the classical path between the end points  $q_j, t''; q_0, t'$ , and is given explicitly in **C**.<sup>7</sup> It is

<sup>7</sup> That (12) is correct, at least insofar as it depends on  $q_0$ , can be seen directly as follows. Let  $\tilde{q}(t)$  be the classical path which satisfies the boundary condition  $\tilde{q}(t') = q_0$ ,  $\tilde{q}(t'') = q_j$ . Then in the integral defining  $k$  replace each of the variables  $q_i$  by  $q_i = \tilde{q}_i + y_i$ , ( $\tilde{q}_i = \tilde{q}(t_i)$ ), that is, use the displacement  $y_i$  from the classical path  $\tilde{q}_i$  as the coordinate rather than the absolute position. With the substitution  $q_i = \tilde{q}_i + y_i$  in the action

$$\begin{aligned} S_0 + S_I &= \int (\frac{1}{2} \dot{q}^2 - \frac{1}{2} \omega^2 q^2 + \gamma q) dt \\ &= \int (\frac{1}{2} \dot{\tilde{q}}^2 - \frac{1}{2} \omega^2 \tilde{q}^2 + \gamma \tilde{q}) dt + \int (\frac{1}{2} \dot{y}^2 - \frac{1}{2} \omega^2 y^2) dt \end{aligned}$$

the terms linear in  $y$  drop out by integrations by parts using the equation of motion  $\ddot{q} = -\omega^2 \tilde{q} + \gamma \tilde{q}$  for the classical path, and the boundary conditions  $y(t') = y(t'') = 0$ . That this should occur should occasion no surprise, for the action functional is an extremum at  $q(t) = \tilde{q}(t)$  so that it will only depend to second order in the displacements  $y$  from this extremal orbit  $\tilde{q}(t)$ . Further, since the action functional is quadratic to begin with, it cannot depend on  $y$  more than quadratically. Hence

$$S_0 + S_I = Q + \int (\frac{1}{2} \dot{y}^2 - \frac{1}{2} \omega^2 y^2) dt$$

so that since  $dq_i = dy_i$ ,

$$k(q_j, t''; q_0, t') = \exp(iQ) \int \exp \left( i \int \frac{1}{2} (y^2 - \omega^2 y^2) dt \right) \mathcal{D}y(t).$$

The factor following the  $\exp(iQ)$  is the amplitude for a free oscillator to proceed from  $y=0$  at  $t=t'$  to  $y=0$  at  $t=t''$  and does not there-

$$Q = \frac{\omega}{2 \sin\omega(t''-t')} \left[ (q_0^2 + q_0^2) \cos\omega(t''-t') - 2q_j q_0 \right. \\ + \frac{2q_j}{\omega} \int_{t'}^{t''} \gamma(t) \sin\omega(t-t') dt \\ + \frac{2q_0}{\omega} \int_{t'}^{t''} \gamma(t) \sin\omega(t''-t) dt \\ - \frac{2}{\omega^2} \int_{t'}^{t''} \int_{t'}^t \gamma(t) \gamma(s) \sin\omega(t''-t) \\ \times \sin\omega(s-t') ds dt \left. \right]. \quad (13)$$

The solution of the motion of the oscillator can now be completed by substituting (12) and (13) into (11) and performing the integrals. The simplest case is for  $m, n=0$  for which case<sup>8</sup>

$$\varphi_0(q_0) = (\omega/\pi)^{\frac{1}{2}} \exp(-\frac{1}{2}\omega q_0^2) \exp(-\frac{1}{2}i\omega t')$$

so that the integrals on  $q_0, q_j$  are just Gaussian integrals. There results

$$G_{00} = \exp\left(-\frac{1}{2}\omega^{-1} \int_{t'}^{t''} \int_{t'}^t \exp(-i\omega(t-s)) \gamma(t) \gamma(s) dt ds\right)$$

a result of fundamental importance in the succeeding developments. By replacing  $t-s$  by its absolute value  $|t-s|$  we may integrate both variables over the entire range and divide by 2. We will henceforth make the results more general by extending the limits on the integrals from  $-\infty$  to  $+\infty$ . Thus if one wishes to study the effect on a particle of interaction with an oscillator for just the period  $t'$  to  $t''$  one may use

$$G_{00} = \exp\left(-\frac{1}{4\omega} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \right. \\ \times \exp(-i\omega|t-s|) \gamma(t) \gamma(s) dt ds \left. \right) \quad (14)$$

imagining in this case that the interaction  $\gamma(t)$  is zero outside these limits. We defer to a later section the discussion of other values of  $m, n$ .

Since  $G_{00}$  is simply an exponential, we can write it as  $\exp(iI)$ , consider that the complete "action" for the system of particles is  $S=S_p+I$  and that one computes transition elements with this "action" instead of  $S_p$ ,

fore depend on  $q_0, q_j$ , or  $\gamma(t)$ , being a function only of  $t''-t'$ . [That it is actually  $(2\pi i\omega^{-1} \sin\omega(t''-t'))^{-1}$  can be demonstrated either by direct integration of the  $y$  variables or by using some normalizing property of the kernels  $k$ , for example that  $G_{00}$  for the case  $\gamma=0$  must equal unity.] The expression for  $Q$  given in C on page 386 is in error, the quantities  $q_0$  and  $q_j$  should be interchanged.

<sup>8</sup> It is most convenient to define the state  $\varphi_n$  with the phase factor  $\exp[-i\omega(n+\frac{1}{2})t']$  and the final state with the factor  $\exp[-i\omega(m+\frac{1}{2})t'']$  so that the results will not depend on the particular times  $t', t''$  chosen.

(see C, Sec. 12). The functional  $I$ , which is given by

$$I = \frac{1}{4}i\omega^{-1} \int \int \exp(-i\omega|t-s|) \gamma(s) \gamma(t) ds dt \quad (15)$$

is complex, however; we shall speak of it as the complex action. It describes the fact that the system at one time can affect itself at a different time by means of a temporary storage of energy in the oscillator. When there are several independent oscillators with different interactions, the effect, if they are all in the lowest state at  $t'$  and  $t''$ , is the product of their separate  $G_{00}$  contributions. Thus the complex action is additive, being the sum of contributions like (15) for each of the several oscillators.

#### 4. VIRTUAL TRANSITIONS IN THE ELECTROMAGNETIC FIELD

We can now apply these results to eliminate the transverse field oscillators of the Lagrangian (2). At first we can limit ourselves to the case of purely virtual transitions in the electromagnetic field, so that there is no photon in the field at  $t'$  and  $t''$ . That is, all of the field oscillators are making transitions from ground state to ground state.

The  $\gamma_{\mathbf{K}^{(r)}}$  corresponding to each oscillator  $q_{\mathbf{K}^{(r)}}$  is found from the interaction term  $L_I$  [Eq. (2b)], substituting the value of  $A^{tr}(\mathbf{x})$  given in (1). There results, for example,

$$\begin{aligned} \gamma_{\mathbf{K}^{(1)}} &= (8\pi)^{\frac{1}{2}} \sum_n e_n (\mathbf{e}_1 \cdot \mathbf{x}_n) \cos(\mathbf{K} \cdot \mathbf{x}_n) \\ \gamma_{\mathbf{K}^{(3)}} &= (8\pi)^{\frac{1}{2}} \sum_n e_n (\mathbf{e}_1 \cdot \mathbf{x}_n) \sin(\mathbf{K} \cdot \mathbf{x}_n) \end{aligned} \quad (16)$$

the corresponding results for  $\gamma_{\mathbf{K}^{(2)}}, \gamma_{\mathbf{K}^{(4)}}$  replace  $\mathbf{e}_1$  by  $\mathbf{e}_2$ .

The complex action resulting from oscillator of coordinate  $q_{\mathbf{K}^{(1)}}$  is therefore

$$I_{\mathbf{K}^{(1)}} = \frac{8\pi i}{4k} \sum_n \sum_m \int \int e_n e_m \exp(-ik|t-s|) (\mathbf{e}_1 \cdot \mathbf{x}_n(t)) \\ \times (\mathbf{e}_1 \cdot \mathbf{x}_m(s)) \cos(\mathbf{K} \cdot \mathbf{x}_n(t)) \cos(\mathbf{K} \cdot \mathbf{x}_m(s)) ds dt.$$

The term  $I_{\mathbf{K}^{(3)}}$  exchanges the cosines for sines, so in the sum  $I_{\mathbf{K}^{(1)}} + I_{\mathbf{K}^{(3)}}$  the product of the two cosines,  $\cos A \cdot \cos B$  is replaced by  $(\cos A \cos B + \sin A \sin B)$  or  $\cos(A-B)$ . The terms  $I_{\mathbf{K}^{(2)}} + I_{\mathbf{K}^{(4)}}$  give the same result with  $\mathbf{e}_2$  replacing  $\mathbf{e}_1$ . The sum  $(\mathbf{e}_1 \cdot \mathbf{V})(\mathbf{e}_1 \cdot \mathbf{V}') + (\mathbf{e}_2 \cdot \mathbf{V})(\mathbf{e}_2 \cdot \mathbf{V}')$  is  $(\mathbf{V} \cdot \mathbf{V}') - k^{-2}(\mathbf{K} \cdot \mathbf{V})(\mathbf{K} \cdot \mathbf{V}')$  since it is the sum of the products of vector components in two orthogonal directions, so that if we add the product in the third direction (that of  $\mathbf{K}$ ) we construct the complete scalar product. Summing over all  $\mathbf{K}$  then, since  $\sum_{\mathbf{K}} = \frac{1}{2} \int d^3 \mathbf{K} / 8\pi^3$  we find for the total complex action of all of the transverse oscillators,

$$I_{tr} = i \sum_n \sum_m \int_{t'}^{t''} dt \int_{t'}^{t''} ds \int e_n e_m \exp(-ik|t-s|) \\ \times [\mathbf{x}_n(t) \cdot \mathbf{x}_m(s) - k^{-2}(\mathbf{K} \cdot \mathbf{x}_n(t))(\mathbf{K} \cdot \mathbf{x}_m(s))] \\ \cdot \cos(\mathbf{K} \cdot (\mathbf{x}_n(t) - \mathbf{x}_m(s))) d^3 \mathbf{K} / 8\pi^3 k. \quad (17)$$

This is to be added to  $S_p + S_c$  to obtain the complete action of the system with the oscillators removed.

The term in  $(\mathbf{K} \cdot \mathbf{x}_n(t))(\mathbf{K} \cdot \mathbf{x}_m(s))$  can be simplified by integration by parts with respect to  $t$  and with respect to  $s$  [note that  $\exp(-ik|t-s|)$  has a discontinuous slope at  $t=s$ , or break the integration up into two regions]. One finds

$$I_{tr} = R - I_c + I_{transient} \quad (18)$$

where

$$\begin{aligned} R = & -i \sum_n \sum_m \int_{t'}^{t''} dt \int_{t'}^{t''} ds \int e_n e_m \\ & \times \exp(-ik|t-s|) (1 - \mathbf{x}_n(t) \cdot \mathbf{x}_m(s)) \\ & \cdot \cos \mathbf{K} \cdot (\mathbf{x}_n(t) - \mathbf{x}_m(s)) d^3 \mathbf{K} / 8\pi^2 k \end{aligned} \quad (19)$$

and

$$\begin{aligned} I_c = & -\sum_n \sum_m \int_{t'}^{t''} dt \int e_n e_m \\ & \times \cos \mathbf{K} \cdot (\mathbf{x}_n(t) - \mathbf{x}_m(t)) d^3 \mathbf{K} / 4\pi^2 k^2 \end{aligned} \quad (20)$$

comes from the discontinuity in slope of  $\exp(-ik|t-s|)$  at  $t=s$ . Since

$$\int \cos(\mathbf{K} \cdot \mathbf{R}) d^3 \mathbf{K} / 4\pi^2 k^2 = \int_0^\infty (kr)^{-1} \sin(kr) dk / \pi = (2r)^{-1}$$

this term  $I_c$  just cancels the Coulomb interaction term  $S_c = \int L_c dt$ . The term

$$\begin{aligned} I_{transient} = & -\sum_n \sum_m e_n e_m \int \frac{d^3 \mathbf{K}}{4\pi^2 k^2} \\ & \times \left\{ \int_{t'}^{t''} [\exp(-ik(t''-t)) \cos \mathbf{K} \cdot (\mathbf{x}_n(t'') - \mathbf{x}_m(t))] dt \right. \\ & + \exp(-ik(t-t')) \cos \mathbf{K} \cdot (\mathbf{x}_n(t) - \mathbf{x}_m(t')) dt \\ & + (2k)^{-1} i [\cos \mathbf{K} \cdot (\mathbf{x}_n(t'') - \mathbf{x}_m(t'')) \\ & + \cos \mathbf{K} \cdot (\mathbf{x}_n(t') - \mathbf{x}_m(t')) \\ & \left. - 2 \exp(-ik(t''-t')) \cos \mathbf{K} \cdot (\mathbf{x}_n(t') - \mathbf{x}_m(t''))] \right\}. \end{aligned} \quad (21)$$

is one which comes from the limits of integration at  $t'$  and  $t''$ , and involves the coordinates of the particle at either one of these times or the other. If  $t'$  and  $t''$  are considered to be exceedingly far in the past and future, there is no correlation to be expected between these temporally distant coordinates and the present ones, so the effects of  $I_{transient}$  will cancel out quantum mechanically by interference. This transient was produced by the sudden turning on of the interaction of field and particles at  $t'$  and its sudden removal at  $t''$ . Alternatively we can imagine the charges to be turned on after  $t'$  adiabatically and turned off slowly before  $t''$  (in this case, in the term  $L_c$ , the charges should also be

considered as varying with time). In this case, in the limit,  $I_{transient}$  is zero.<sup>9</sup> Hereafter we shall drop the transient term and consider the range of integration of  $t$  to be from  $-\infty$  to  $+\infty$ , imagining, if one needs a definition, that the charges vary with time and vanish in the direction of either limit.

To simplify  $R$  we need the integral

$$\begin{aligned} J = & \int \exp(-ik|t|) \cos(\mathbf{K} \cdot \mathbf{R}) d^3 \mathbf{K} / 8\pi^2 k \\ = & \int_0^\infty \exp(-ik|t|) \sin(kr) dk / 2\pi r \end{aligned} \quad (22)$$

where  $r$  is the length of the vector  $\mathbf{R}$ . Now

$$\begin{aligned} \int_0^\infty \exp(-ikx) dk = & \lim_{\epsilon \rightarrow 0} (-i(x-i\epsilon)^{-1}) \\ = & -ix^{-1} + \pi\delta(x) = \pi\delta_+(x) \end{aligned}$$

where the equation serves to define  $\delta_+(x)$  [as in II, Eq. (3)]. Hence, expanding  $\sin(kr)$  in exponentials find

$$\begin{aligned} J = & -(4\pi r)^{-1} ((|t|-r)^{-1} - (|t|+r)^{-1}) \\ & + (4ir)^{-1} (\delta(|t|-r) - \delta(|t|+r)) \\ = & -(2\pi)^{-1} (t^2 - r^2)^{-1} + (2i)^{-1} \delta(t^2 - r^2) \\ = & -\frac{1}{2} i \delta_+(t^2 - r^2) \end{aligned} \quad (23)$$

where we have used the fact that

$$\delta(t^2 - r^2) = (2r)^{-1} (\delta(|t|-r) + \delta(|t|+r))$$

and that  $\delta(|t|+r)=0$  since both  $|t|$  and  $r$  are necessarily positive.

Substitution of these results into (19) gives finally,

$$\begin{aligned} R = & -\frac{1}{2} \sum_n \sum_m \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e_n e_m (1 - \mathbf{x}_n(t) \cdot \mathbf{x}_m(s)) \\ & \times \delta_+((t-s)^2 - (\mathbf{x}_n(t) - \mathbf{x}_m(s))^2) dt ds. \end{aligned} \quad (24)$$

The total complex action of the system is then<sup>10</sup>  $S_p + R$ . Or, what amounts to the same thing; to obtain

<sup>9</sup> One can obtain the final result, that the total interaction is just  $R$ , in a formal manner starting from the Hamiltonian from which the longitudinal oscillators have not yet been eliminated. There are for each  $\mathbf{K}$  and cos or sin, four oscillators  $q_{\mu\mathbf{K}}$  corresponding to the three components of the vector potential ( $\mu=1, 2, 3$ ) and the scalar potential ( $\mu=4$ ). It must then be assumed that the wave functions of the initial and final state of the  $\mathbf{K}$  oscillators is the function  $(k/\pi) \exp[-\frac{1}{2}k(q_{1\mathbf{K}}^2 + q_{2\mathbf{K}}^2 + q_{3\mathbf{K}}^2 - q_{4\mathbf{K}}^2)]$ . The wave function suggested here has only formal significance, of course, because the dependence on  $q_{4\mathbf{K}}$  is not square integrable, and cannot be normalized. If each oscillator were assumed actually in the ground state, the sign of the  $q_{4\mathbf{K}}$  term would be changed to positive, and the sign of the frequency in the contribution of these oscillators would be reversed (they would have negative energy).

<sup>10</sup> The classical action for this problem is just  $S_p + R'$  where  $R'$  is the real part of the expression (24). In view of the generalization of the Lagrangian formulation of quantum mechanics suggested in Section 12 of C, one might have anticipated that  $R$  would have been simply  $R'$ . This corresponds, however, to boundary conditions other than no quanta present in past and future. It is harder to interpret physically. For a system enclosed in a light tight box, however, it appears likely that both  $R$  and  $R'$  lead to the same results.

transition amplitudes including the effects of the field we must calculate the transition element of  $\exp(iR)$ :

$$\langle \chi_{t'} | \exp iR | \psi_{t'} \rangle_{S_p} \quad (25)$$

under the action  $S_p$  of the particles, excluding interaction. Expression (24) for  $R$  must be considered to be written in the usual manner as a Riemann sum and the expression (25) interpreted as defined in **C** [Eq. (39)]. Expression (6) must be used for  $\mathbf{x}_n$  at time  $t$ .

Expression (25), with (24), then contains all the effects of virtual quanta on a (at least non-relativistic) system according to quantum electrodynamics. It contains the effects to all orders in  $e^2/\hbar c$  in a single expression. If expanded in a power series in  $e^2/\hbar c$ , the various terms give the expressions to the corresponding order obtained by the diagrams and methods of **II**. We illustrate this by an example in the next section.

##### 5. EXAMPLE OF APPLICATION OF EXPRESSION (25)

We shall not be much concerned with the non-relativistic case here, as the relativistic case given below is as simple and more interesting. It is, however, very similar and at this stage it is worth giving an example to show how expressions resulting from (25) are to be interpreted according to the rules of **C**. For example, consider the case of a single electron, coordinate  $\mathbf{x}$ , either free or in an external given potential (contained for simplicity in  $S_p$ , not in<sup>11</sup>  $R$ ). Its interaction with the field produces a reaction back on itself given by  $R$  as in (24) but in which we keep only a single term corresponding to  $m=n$ . Assume the effect of  $R$  to be small and expand  $\exp(iR)$  as  $1+iR$ . Let us find the amplitude at time  $t''$  of finding the electron in a state  $\psi$  with no quanta emitted, if at time  $t'$  it was in the same state. It is

$$\langle \psi_{t'} | 1 + iR | \psi_{t'} \rangle_{S_p} = \langle \psi_{t'} | 1 | \psi_{t'} \rangle_{S_p} + i \langle \psi_{t'} | R | \psi_{t'} \rangle_{S_p}$$

where  $\langle \psi_{t'} | 1 | \psi_{t'} \rangle_{S_p} = \exp[-iE(t''-t')]$  if  $E$  is the energy of the state, and

$$\begin{aligned} \langle \psi_{t'} | R | \psi_{t'} \rangle_{S_p} &= -\frac{1}{2} e^2 \int_{t'}^{t''} dt \int_s^{t''} ds \langle \psi_{t'} | (1 - \mathbf{x}_{t'} \cdot \mathbf{x}_s) \\ &\quad \times \delta_+((t-s)^2 - (\mathbf{x}_t - \mathbf{x}_s)^2) | \psi_t \rangle_{S_p}. \end{aligned} \quad (26)$$

Here  $\mathbf{x}_s = \mathbf{x}(s)$ , etc. In (26) we shall limit the range of integrations by assuming  $s < t$ , and double the result.

The expression within the brackets  $\langle \dots \rangle_{S_p}$  on the right-hand side of (26) can be evaluated by the methods described in **C** [Eq. (29)]. An expression such as (26)

<sup>11</sup> One can show from (25) how the correlated effect of many atoms at a distance produces on a given system the effects of an external potential. Formula (24) yields the result that this potential is that obtained from Liénard and Wiechert by retarded waves arising from the charges and currents resulting from the distant atoms making transitions. Assume the wave functions  $\chi$  and  $\psi$  can be split into products of wave functions for system and distant atoms and expand  $\exp(iR)$  assuming the effect of any individual distant atom is small. Coulomb potentials arise even from nearby particles if they are moving slowly.

can also be evaluated directly in terms of the propagation kernel  $K(2, 1)$  [see **I**, Eq. (2)] for an electron moving in the given potential.

The term  $\mathbf{x}_s \cdot \mathbf{x}_{t'}$  in the non-relativistic case produces an interesting complication which does not have an analog for the relativistic case with the Dirac equation. We discuss it below, but for a moment consider in further detail expression (26) but with the factor  $(1 - \mathbf{x}_s \cdot \mathbf{x}_{t'})$  replaced simply by unity.

The kernel  $K(2, 1)$  is defined and discussed in **I**. From its definition as the amplitude that the electron be found at  $\mathbf{x}_2$  at time  $t_2$ , if at  $t_1$  it was at  $\mathbf{x}_1$ , we have

$$K(\mathbf{x}_2, t_2; \mathbf{x}_1, t_1) = \langle \delta(\mathbf{x} - \mathbf{x}_2)_{t_2} | \delta(\mathbf{x} - \mathbf{x}_1)_{t_1} \rangle_{S_p} \quad (27)$$

that is, more simply  $K(2, 1)$  is the sum of  $\exp(iS_p)$  over all paths which go from space time point 1 to 2.

In the integrations over all paths implied by the symbol in (26) we can first integrate over all the  $\mathbf{x}_i$  variables corresponding to times  $t_i$  from  $t'$  to  $s$ , not inclusive, the result being a factor  $K(\mathbf{x}_s, s; \mathbf{x}_{t'}, t')$  according to (27). Next we integrate on the variables between  $s$  and  $t$  not inclusive, giving a factor  $K(\mathbf{x}_t, t; \mathbf{x}_s, s)$  and finally on those between  $t$  and  $t''$  giving  $K(\mathbf{x}_{t'}, t''; \mathbf{x}_t, t)$ . Hence the left-hand term in (26) excluding the  $\mathbf{x}_{t'} \cdot \mathbf{x}_s$  factor is

$$\begin{aligned} -e^2 \int dt \int ds \int \psi^*(\mathbf{x}_{t'}, t') K(\mathbf{x}_{t'}, t'; \mathbf{x}_t, t) \delta_+((t-s)^2 \\ - (\mathbf{x}_t - \mathbf{x}_s)^2) \cdot K(\mathbf{x}_t, t; \mathbf{x}_s, s) K(\mathbf{x}_s, s; \mathbf{x}_{t'}, t') \\ \times \psi(\mathbf{x}_{t'}, t') d^3 \mathbf{x}_{t'} d^3 \mathbf{x}_t d^3 \mathbf{x}_s d^3 \mathbf{x}_{t'} \end{aligned} \quad (28)$$

which in improved notation and in the relativistic case is essentially the result given in **II**.

We have made use of a special case of a principle which may be stated more generally as

$$\begin{aligned} \langle \chi_{t'} | F(\mathbf{x}_1, t_1; \mathbf{x}_2, t_2; \dots; \mathbf{x}_k, t_k) | \psi_{t'} \rangle_{S_p} \\ = \int \chi^*(\mathbf{x}_{t'}) K(\mathbf{x}_{t'}, t'; \mathbf{x}_1, t_1) \cdot K(\mathbf{x}_1, t_1; \mathbf{x}_2, t_2) \dots \\ \times K(\mathbf{x}_{k-1}, t_{k-1}; \mathbf{x}_k, t_k) K(\mathbf{x}_k, t_k; \mathbf{x}_{t'}, t') \\ \cdot F(\mathbf{x}_1, t_1; \mathbf{x}_2, t_2; \dots; \mathbf{x}_k, t_k) \psi(\mathbf{x}_{t'}) \\ \times d^3 \mathbf{x}_{t'} d^3 \mathbf{x}_1 d^3 \mathbf{x}_2 \dots d^3 \mathbf{x}_k d^3 \mathbf{x}_{t'} \end{aligned} \quad (29)$$

where  $F$  is any function of the coordinate  $\mathbf{x}_1$  at time  $t_1$ ,  $\mathbf{x}_2$  at  $t_2$  up to  $\mathbf{x}_k$ ,  $t_k$ , and, it is important to notice, we have assumed  $t'' > t_1 > t_2 > \dots > t_k > t'$ .

Expressions of higher order arising for example from  $R^2$  are more complicated as there are quantities referring to several different times mixed up, but they all can be interpreted readily. One simply breaks up the ranges of integrations of the time variables into parts such that in each the order of time of each variable is definite. One then interprets each part by formula (29).

As a simple example we may refer to the problem of the transition element

$$\left\langle \chi_{t''} \left| \int U(\mathbf{x}(t), t) dt \int V(\mathbf{x}(s), s) ds \right| \psi_{t'} \right\rangle$$

arising, say, in the cross term in  $U$  and  $V$  in an ordinary second order perturbation problem (disregarding radiation) with perturbation potential  $U(\mathbf{x}, t) + V(\mathbf{x}, t)$ . In the integration on  $s$  and  $t$  which should include the entire range of time for each, we can split the range of  $s$  into two parts,  $s < t$  and  $s > t$ . In the first case,  $s < t$ , the potential  $V$  acts earlier than  $U$ , and in the other range, vice versa, so that

$$\begin{aligned} & \left\langle \chi_{t''} \left| \int U(\mathbf{x}_t, t) dt \int V(\mathbf{x}_s, s) ds \right| \psi_{t'} \right\rangle \\ &= \int_{t'}^{t''} dt \int_t^t ds \int \chi^*(\mathbf{x}_{t''}) K(\mathbf{x}_{t''}, t''; \mathbf{x}_t, t) \\ & \quad \times U(\mathbf{x}_t, t) K(\mathbf{x}_t, t; \mathbf{x}_s, s) V(\mathbf{x}_s, s) \\ & \quad \cdot K(\mathbf{x}_s, s; \mathbf{x}_{t'}, t') \psi(\mathbf{x}_{t'}) d^3 \mathbf{x}_{t'} d^3 \mathbf{x}_t d^3 \mathbf{x}_{t''} \\ &+ \int_{t'}^{t''} dt \int_t^{t''} ds \int \chi^*(\mathbf{x}_{t''}) K(\mathbf{x}_{t''}, t''; \mathbf{x}_s, s) \\ & \quad \times V(\mathbf{x}_s, s) K(\mathbf{x}_s, s; \mathbf{x}_t, t) U(\mathbf{x}_t, t) \\ & \quad \cdot K(\mathbf{x}_t, t; \mathbf{x}_{t'}, t') \psi(\mathbf{x}_{t'}) d^3 \mathbf{x}_{t'} d^3 \mathbf{x}_s d^3 \mathbf{x}_t d^3 \mathbf{x}_{t''} \end{aligned} \quad (30)$$

so that the single expression on the left is represented by two terms analogous to the two terms required in analyzing the Compton effect. It is in this way that the several terms and their corresponding diagrams corresponding to each process arise when an attempt is made to represent the transition elements of single expressions involving time integrals in terms of the propagation kernels  $K$ .

It remains to study in more detail the term in (26) arising from  $\mathbf{x}'(t) \cdot \mathbf{x}'(s)$  in the interaction. The interpretation of such expressions is considered in detail in C, and we must refer to Eqs. (39) through (50) of that paper for a more thorough analysis. A similar type of term also arises in the Lagrangian formulation in simpler problems, for example the transition element

$$\left\langle \chi_{t''} \left| \int \mathbf{x}'(t) \cdot \mathbf{A}(\mathbf{x}(t), t) dt \int \mathbf{x}'(s) \cdot \mathbf{B}(\mathbf{x}(s), s) ds \right| \psi_{t'} \right\rangle$$

arising say, in the cross term in  $\mathbf{A}$  and  $\mathbf{B}$  in a second-order perturbation problem for a particle in a perturbing vector potential  $\mathbf{A}(\mathbf{x}, t) + \mathbf{B}(\mathbf{x}, t)$ . The time integrals must first be written as Riemannian sums, the velocity (see (6)) being replaced by  $\mathbf{x}' = \frac{1}{2}\epsilon^{-1}(\mathbf{x}_{i+1} - \mathbf{x}_i) + \frac{1}{2}\epsilon^{-1}(\mathbf{x}_i - \mathbf{x}_{i-1})$  so that we ask for the transition

element of

$$\sum_i \sum_j [\frac{1}{2}(\mathbf{x}_{i+1} - \mathbf{x}_i) + \frac{1}{2}(\mathbf{x}_i - \mathbf{x}_{i-1})] \cdot \mathbf{A}(\mathbf{x}_i, t_i) \\ \times [\frac{1}{2}(\mathbf{x}_{j+1} - \mathbf{x}_j) + \frac{1}{2}(\mathbf{x}_j - \mathbf{x}_{j-1})] \cdot \mathbf{B}(\mathbf{x}_j, t_j). \quad (31)$$

In C it is shown that when converted to operator notation the quantity  $(\mathbf{x}_{i-1} - \mathbf{x}_i)/\epsilon$  is equivalent (nearly, see below) to an operator,

$$(\mathbf{x}_{i+1} - \mathbf{x}_i)/\epsilon \rightarrow i(H\mathbf{x} - \mathbf{x}H) \quad (32)$$

operating in order indicated by the time index  $i$  (that is after  $x_l$ 's for  $l \leq i$  and before all  $x_l$ 's for  $l > i$ ). In non-relativistic mechanics  $i(H\mathbf{x} - \mathbf{x}H)$  is the momentum operator  $p_x$  divided by the mass  $m$ . Thus in (31) the expression  $[\frac{1}{2}(\mathbf{x}_{i+1} - \mathbf{x}_i) + \frac{1}{2}(\mathbf{x}_i - \mathbf{x}_{i-1})] \cdot \mathbf{A}(\mathbf{x}_i, t_i)$  becomes  $\epsilon(\mathbf{p} \cdot \mathbf{A} + \mathbf{A} \cdot \mathbf{p})/2m$ . Here again we must split the sum into two regions  $j < i$  and  $j > i$  so the quantities in the usual notation will operate in the right order such that eventually (31) becomes identical with the right-hand side of Eq. (30) but with  $U(\mathbf{x}_t, t)$  replaced by the operator

$$\frac{1}{2m} \left( \frac{1}{i} \frac{\partial}{\partial \mathbf{x}_t} \cdot \mathbf{A}(\mathbf{x}_t, t) + \mathbf{A}(\mathbf{x}_t, t) \cdot \frac{1}{i} \frac{\partial}{\partial \mathbf{x}_t} \right)$$

standing in the same place, and with the operator

$$\frac{1}{2m} \left( \frac{1}{i} \frac{\partial}{\partial \mathbf{x}_s} \cdot \mathbf{B}(\mathbf{x}_s, s) + \frac{1}{i} \mathbf{B}(\mathbf{x}_s, s) \cdot \frac{\partial}{\partial \mathbf{x}_s} \right)$$

standing in the place of  $V(\mathbf{x}_s, s)$ . The sums and factors  $\epsilon$  have now become  $\int dt \int ds$ .

This is nearly but not quite correct, however, as there is an additional term coming from the terms in the sum corresponding to the special values,  $j=i$ ,  $j=i+1$  and  $j=i-1$ . We have tacitly assumed from the appearance of the expression (31) that, for a given  $i$ , the contribution from just three such special terms is of order  $\epsilon^2$ . But this is not true. Although the expected contribution of a term like  $(\mathbf{x}_{i+1} - \mathbf{x}_i)(\mathbf{x}_{j+1} - \mathbf{x}_j)$  for  $j \neq i$  is indeed of order  $\epsilon^2$ , the expected contribution of  $(\mathbf{x}_{i+1} - \mathbf{x}_i)^2$  is  $+i\epsilon m^{-1}$  [C, Eq. (50)], that is, of order  $\epsilon$ . In non-relativistic mechanics the velocities are unlimited and in very short times  $\epsilon$  the amplitude diffuses a distance proportional to the square root of the time. Making use of this equation then we see that the additional contribution from these terms is essentially

$$im^{-1} \epsilon \sum_i \mathbf{A}(\mathbf{x}_i, t_i) \cdot \mathbf{B}(\mathbf{x}_i, t_i) = im^{-1} \int \mathbf{A}(\mathbf{x}(t), t) \cdot \mathbf{B}(\mathbf{x}(t), t) dt$$

when summed on all  $i$ . This has the same effect as a first-order perturbation due to a potential  $\mathbf{A} \cdot \mathbf{B}/m$ . Added to the term involving the momentum operators

we therefore have an additional term<sup>12</sup>

$$\frac{i}{m} \int_{t'}^{t''} dt \int \chi^*(\mathbf{x}_{t''}) K(\mathbf{x}_{t''}, t'; \mathbf{x}_t, t) \mathbf{A}(\mathbf{x}_t, t) \cdot \mathbf{B}(\mathbf{x}_t, t) \cdot K(\mathbf{x}_t, t; \mathbf{x}_{t'}, t') \psi(\mathbf{x}_{t'}) d^3\mathbf{x}_{t''} d^3\mathbf{x}_t d^3\mathbf{x}_{t'}. \quad (33)$$

In the usual Hamiltonian theory this term arises, of course, from the term  $\mathbf{A}^2/2m$  in the expansion of the Hamiltonian

$$H = (2m)^{-1}(\mathbf{p} - \mathbf{A})^2 = (2m)^{-1}(\mathbf{p}^2 - \mathbf{p} \cdot \mathbf{A} - \mathbf{A} \cdot \mathbf{p} + \mathbf{A}^2)$$

while the other term arises from the second-order action of  $\mathbf{p} \cdot \mathbf{A} + \mathbf{A} \cdot \mathbf{p}$ . We shall not be interested in non-relativistic quantum electrodynamics in detail. The situation is simpler for Dirac electrons. For particles satisfying the Klein-Gordon equation (discussed in Appendix A) the situation is very similar to a four-dimensional analog of the non-relativistic case given here.

## 6. EXTENSION TO DIRAC PARTICLES

Expressions (24) and (25) and their proof can be readily generalized to the relativistic case according to the one electron theory of Dirac. We shall discuss the hole theory later. In the non-relativistic case we began with the proposition that the amplitude for a particle to proceed from one point to another is the sum over paths of  $\exp(iS_p)$ , that is, we have for example for a transition element

$$\langle \mathbf{x}_1 | \psi \rangle = \lim_{\epsilon \rightarrow 0} \int \cdots \int \chi^*(\mathbf{x}_N) \Phi_p(\mathbf{x}_N, \mathbf{x}_{N-1}, \dots, \mathbf{x}_0) \cdot \psi(\mathbf{x}_0) d^3\mathbf{x}_0 d^3\mathbf{x}_1 \cdots d^3\mathbf{x}_N \quad (34)$$

where for  $\exp(iS_p)$  we have written  $\Phi_p$ , that is more precisely,

$$\Phi_p = \Pi_i A^{-1} \exp[iS(\mathbf{x}_{i+1}, \mathbf{x}_i)].$$

As discussed in C this form is related to the usual form of quantum mechanics through the observation that

$$(\mathbf{x}_{i+1} | \mathbf{x}_i)_\epsilon = A^{-1} \exp[iS(\mathbf{x}_{i+1}, \mathbf{x}_i)] \quad (35)$$

where  $(\mathbf{x}_{i+1} | \mathbf{x}_i)_\epsilon$  is the transformation matrix from a representation in which  $\mathbf{x}$  is diagonal at time  $t_i$  to one in which  $\mathbf{x}$  is diagonal at time  $t_{i+1} = t_i + \epsilon$  (so that it is identical to  $K_0(\mathbf{x}_{i+1}, t_{i+1}; \mathbf{x}_i, t_i)$  for the small time interval  $\epsilon$ ). Hence the amplitude for a given path can also be written

$$\Phi_p = \Pi_i (\mathbf{x}_{i+1} | \mathbf{x}_i)_\epsilon \quad (36)$$

for which form, of course, (34) is exact irrespective of whether  $(\mathbf{x}_{i+1} | \mathbf{x}_i)_\epsilon$  can be expressed in the simple form (35).

For a Dirac electron the  $(\mathbf{x}_{i+1} | \mathbf{x}_i)_\epsilon$  is a  $4 \times 4$  matrix

<sup>12</sup> The term corresponding to this for the self-energy expression (26) would give an integral over  $\delta_+(\!(t-t')^2 - (\mathbf{x}_t - \mathbf{x}_{t'})^2)$  which is evidently infinite and leads to the quadratically divergent self-energy. There is no such term for the Dirac electron, but there is for Klein-Gordon particles. We shall not discuss the infinities in this paper as they have already been discussed in II.

(or  $4^N \times 4^N$  if we deal with  $N$  electrons) but the expression (34) with (36) is still correct (as it is in fact for any quantum-mechanical system with a sufficiently general definition of the coordinate  $\mathbf{x}_i$ ). The product (36) now involves operators, the order in which the factors are to be taken is the order in which the terms appear in time.

For a Dirac particle in a vector and scalar potential (times the electron charge  $e$ )  $\mathbf{A}(\mathbf{x}, t)$ ,  $A_4(\mathbf{x}, t)$ , the quantity  $(\mathbf{x}_{i+1} | \mathbf{x}_i)_\epsilon^{(A)}$  is related to that of a free particle to the first order in  $\epsilon$  as

$$(\mathbf{x}_{i+1} | \mathbf{x}_i)_\epsilon^{(A)} = (\mathbf{x}_{i+1} | \mathbf{x}_i)_\epsilon^{(0)} \exp[-i(\epsilon A_4(\mathbf{x}_i, t_i) - (\mathbf{x}_{i+1} - \mathbf{x}_i) \cdot \mathbf{A}(\mathbf{x}_i, t_i))]. \quad (37)$$

This can be verified directly by substitution into the Dirac equation.<sup>13</sup> It neglects the variation of  $\mathbf{A}$  and  $A_4$  with time and space during the short interval  $\epsilon$ . This produces errors only of order  $\epsilon^2$  in the Dirac case for the expected square velocity  $(\mathbf{x}_{i+1} - \mathbf{x}_i)^2/\epsilon^2$  during the interval  $\epsilon$  is finite (equaling the square of the velocity of light) rather than being of order  $1/\epsilon$  as in the non-relativistic case. [This makes the relativistic case somewhat simpler in that it is not necessary to define the velocity as carefully as in (6);  $(\mathbf{x}_{i+1} - \mathbf{x}_i)/\epsilon$  is sufficiently exact, and no term analogous to (33) arises.]

Thus  $\Phi_p^{(A)}$  differs from that for a free particle,  $\Phi_p^{(0)}$ , by a factor  $\Pi_i \exp[-i(\epsilon A_4(\mathbf{x}_i, t_i) - (\mathbf{x}_{i+1} - \mathbf{x}_i) \cdot \mathbf{A}(\mathbf{x}_i, t_i))]$  which in the limit can be written as

$$\exp \left\{ -i \int [A_4(\mathbf{x}(t), t) - \mathbf{x}(t) \cdot \mathbf{A}(\mathbf{x}(t), t)] dt \right\} \quad (38)$$

exactly as in the non-relativistic case.

The case of a Dirac particle interacting with the quantum-mechanical oscillators representing the field may now be studied. Since the dependence of  $\Phi_p^{(A)}$  on  $\mathbf{A}$ ,  $A_4$  is through the same factor as in the non-relativistic case, when  $\mathbf{A}$ ,  $A_4$  are expressed in terms of the oscillator coordinates  $q$ , the dependence of  $\Phi$  on the oscillator coordinates  $q$  is unchanged. Hence the entire analysis of the preceding sections which concern the results of the integration over oscillator coordinates can be carried through unchanged and the results will be expression (25) with formula (24) for  $R$ . Expression (25) is now interpreted as

$$\begin{aligned} \langle \mathbf{x}_{t''} | \exp iR | \psi_{t'} \rangle &= \lim_{\epsilon \rightarrow 0} \int \chi^*(\mathbf{x}_{t''}^{(1)}, \mathbf{x}_{t''}^{(2)} \dots) \\ &\quad \times \prod_n (\Phi_p)_n^{(0)} d^3\mathbf{x}_{t''}^{(n)} d^3\mathbf{x}_{t''}^{(n)} \dots d^3\mathbf{x}_{t''}^{(n)} \\ &\quad \times \exp(iR) \psi(\mathbf{x}_{t'}^{(1)}, \mathbf{x}_{t'}^{(2)} \dots) \end{aligned} \quad (39)$$

<sup>13</sup> Alternatively, note that Eq. (37) is exact for arbitrarily large  $\epsilon$  if the potential  $A_\mu$  is constant. For if the potential in the Dirac equation is the gradient of a scalar function  $A_\mu = \partial \chi / \partial x_\mu$  the potential may be removed by replacing the wave function by  $\psi = \exp(-i\chi)\psi'$  (gauge transformation). This alters the kernel by a factor  $\exp[-i(\chi(2) - \chi(1))]$  owing to the change in the initial and final wave functions. A constant potential  $A_\mu$  is the gradient of  $\chi = A_\mu x_\mu$  and can be completely removed by this gauge transformation, so that the kernel differs from that of a free particle by the factor  $\exp[-i(A_\mu x_{\mu 2} - A_\mu x_{\mu 1})]$  as in (37).

where  $\Phi_{p,n}^{(0)}$ , the amplitude for a particular path for particle  $n$  is simply the expression (36) where  $(\mathbf{x}_{i+1}|\mathbf{x}_i)_\epsilon$  is the kernel  $K_{0,n}(\mathbf{x}_{i+1}^{(n)}, t_{i+1}; \mathbf{x}_i^{(n)}, t_i)$  for a free electron according to the one electron Dirac theory, with the matrices which appear operating on the spinor indices corresponding to particle  $(n)$  and the order of all operations being determined by the time indices.

For calculational purposes we can, as before, expand  $R$  as a power series and evaluate the various terms in the same manner as for the non-relativistic case. In such an expansion the quantity  $\mathbf{x}'(t)$  is replaced, as we have seen in (32), by the operator  $i(H\mathbf{x}-\mathbf{x}H)$ , that is, in this case by  $\alpha$  operating at the corresponding time. There is no further complicated term analogous to (33) arising in this case, for the expected value of  $(x_{i+1}-x_i)^2$  is now of order  $e^2$  rather than  $\epsilon$ .

For example, for self-energy one sees that expression (28) will be (with other terms coming from those with  $\mathbf{x}'(t)$  replaced by  $\alpha$  and with the usual  $\beta$  in back of each  $K_0$  because of the definition of  $K_0$  in relativity theory)

$$\begin{aligned} \langle \psi_{\nu'} | R | \psi_\nu \rangle s_p = & -e^2 \int \psi^*(\mathbf{x}_{\nu'}) K_0(\mathbf{x}_{\nu'}, t''; \mathbf{x}_t, t) \beta \alpha_\mu \\ & \cdot \delta_+((t-s)^2 - (\mathbf{x}_t - \mathbf{x}_s)^2) K_0(\mathbf{x}_t, t; \mathbf{x}_s, s) \beta \alpha_\mu \\ & \cdot K_0(\mathbf{x}_s, s; \mathbf{x}_{\nu'}, t') \beta \psi(\mathbf{x}_{\nu'}) d^3 \mathbf{x}_{\nu'} d^3 \mathbf{x}_t d^3 \mathbf{x}_s d^3 \mathbf{x}_{\nu'} dt ds, \end{aligned} \quad (40)$$

where  $\alpha_4=1$ ,  $\alpha_{1,2,3}=\alpha_{x,y,z}$  and a sum on the repeated index  $\mu$  is implied in the usual way;  $a_\mu b_\mu = a_4 b_4 - a_1 b_1 - a_2 b_2 - a_3 b_3$ . One can change  $\beta \alpha_\mu$  to  $\gamma_\mu$  and  $\psi^*$  to  $\bar{\psi} \beta$ . In this manner all of the rules referring to virtual photons discussed in II are deduced; but with the difference that  $K_0$  is used instead of  $K_+$  and we have the Dirac one electron theory with negative energy states (although we may have any number of such electrons).

## 7. EXTENSION TO POSITRON THEORY

Since in (39) we have an arbitrary number of electrons, we can deal with the hole theory in the usual manner by imagining that we have an infinite number of electrons in negative energy states.

On the other hand, in paper I on the theory of positrons, it was shown that the results of the hole theory in a system with a given external potential  $A_\mu$  were equivalent to those of the Dirac one electron theory if one replaced the propagation kernel,  $K_0$ , by a different one,  $K_+$ , and multiplied the resultant amplitude by factor  $C_v$  involving  $A_\mu$ . We must now see how this relation, derived in the case of external potentials, can also be carried over in electrodynamics to be useful in simplifying expressions involving the infinite sea of electrons.

To do this we study in greater detail the relation between a problem involving virtual photons and one involving purely external potentials. In using (25) we shall assume in accordance with the hole theory that

the number of electrons is infinite, but that they all have the same charge,  $e$ . Let the states  $\psi_\nu, \chi_{\nu'}$ , represent the vacuum plus perhaps a number of real electrons in positive energy states and perhaps also some empty negative energy states. Let us call the amplitude for the transition in an external potential  $B_\mu$ , but *excluding virtual photons*,  $T_0[B]$ , a functional of  $B_\mu(1)$ . We have seen (38)

$$T_0[B] = \langle \chi_{\nu'} | \exp(iP) | \psi_\nu \rangle \quad (41)$$

where

$$P = -\sum_n \int [B_4(\mathbf{x}^{(n)}(t), t) - \mathbf{x}^{(n)}(t) \cdot \mathbf{B}(\mathbf{x}^{(n)}(t), t)] dt$$

by (38). We can write this as

$$P = -\sum_n \int B_\mu(x_\nu^{(n)}(t)) \dot{x}_\mu^{(n)}(t) dt$$

where  $x_4(t)=t$  and  $\dot{x}_4=1$ , the other values of  $\mu$  corresponding to space variables. The corresponding amplitude for the same process in the same potential, but *including* all the virtual photons we may call,

$$T_{e^2}[B] = \langle \chi_{\nu'} | \exp(iR) \exp(ip) | \psi_\nu \rangle. \quad (42)$$

Now let us consider the effect on  $T_{e^2}[B]$  of changing the coupling  $e^2$  of the virtual photons. Differentiating (42) with respect to  $e^2$  which appears only<sup>14</sup> in  $R$  we find

$$\begin{aligned} dT_{e^2}[B]/d(e^2) = & \left\langle \chi_{\nu'} \left| -\frac{i}{2} \sum_n \sum_m \int \int dt ds \dot{x}_\mu^{(n)}(t) \dot{x}_\mu^{(m)}(s) \right. \right. \\ & \cdot \delta_+((x_\nu^{(n)}(t) - x_\nu^{(m)}(s))^2) \exp(i(R+P)) \left. \right| \psi_\nu \rangle. \end{aligned} \quad (43)$$

We can also study the first-order effect of a change of  $B_\mu$ :

$$\begin{aligned} \delta T_{e^2}[B]/\delta B_\mu(1) = & -i \left\langle \chi_{\nu'} \left| \sum_n \int dt \dot{x}_\mu^{(n)} \delta^4(x_\alpha^{(n)}(t) - x_{\alpha,1}) \right. \right. \\ & \cdot \exp(i(R+P)) \left. \right| \psi_\nu \rangle \end{aligned} \quad (44)$$

where  $x_{\alpha,1}$  is the field point at which the derivative with respect to  $B_\mu$  is taken<sup>15</sup> and the term (current density)  $-\sum_n \int dt \dot{x}_\mu^{(n)}(t) \delta^4(x_\alpha^{(n)}(t) - x_{\alpha,1})$  is just  $\delta P/\delta B_\mu(1)$ . The function  $\delta^4(x_\alpha^{(n)} - x_{\alpha,1})$  means  $\delta(x_4^{(n)} - x_{4,1})$

<sup>14</sup> In changing the charge  $e^2$  we mean to vary only the degree to which virtual photons are important. We do not contemplate changes in the influence of the external potentials. If one wishes, as  $e$  is raised the strength of the potential is decreased proportionally so that  $B_\mu$ , the potential times the charge  $e$ , is held constant.

<sup>15</sup> The functional derivative is defined such that if  $T[B]$  is a number depending on the functions  $B_\mu(1)$ , the first order variation in  $T$  produced by a change from  $B_\mu$  to  $B_\mu + \Delta B_\mu$  is given by

$$T[B+\Delta B] - T[B] = \int (\delta T[B]/\delta B_\mu(1)) \Delta B_\mu(1) d\tau_1$$

the integral extending over all four-space  $x_{\alpha,1}$ .

$\times \delta(x_3^{(n)} - x_{3,1})\delta(x_2^{(n)} - x_{2,1})\delta(x_1^{(n)} - x_{1,1})$  that is,  $\delta(2, 1)$  with  $x_{\alpha,2} = x_{\alpha}^{(n)}(t)$ . A second variation of  $T$  gives, by differentiation of (44) with respect to  $B_{\nu}(2)$ ,

$$\begin{aligned} \delta^2 T_{e2}[B]/\delta B_{\mu}(1)\delta B_{\nu}(2) \\ = - \left\langle \chi_{\nu'} \left| \sum_n \sum_m \int dt \int ds \dot{x}_{\mu}^{(n)}(t) \dot{x}_{\nu}^{(m)}(s) \right. \right. \\ \cdot \delta^4(x_{\alpha}^{(n)}(t) - x_{\alpha,1}) \delta^4(x_{\beta}^{(n)}(s) - x_{\beta,2}) \\ \left. \left. \times \exp(i(R+P)) \psi_{\nu'} \right\rangle \right. \end{aligned}$$

Comparison of this with (43) shows that

$$\begin{aligned} dT_{e2}[B]/d(e^2) = \frac{1}{2} i \int \int (\delta^2 T_{e2}[B]/\delta B_{\mu}(1)\delta B_{\mu}(2)) \\ \times \delta_{+}(s_{12}^2) d\tau_1 d\tau_2 \quad (45) \end{aligned}$$

where  $s_{12}^2 = (x_{\mu,1} - x_{\mu,2})(x_{\mu,1} - x_{\mu,2})$ .

We now proceed to use this equation to prove the validity of the rules given in **II** for electrodynamics. This we do by the following argument. The equation can be looked upon as a differential equation for  $T_{e2}[B]$ . It determines  $T_{e2}[B]$  uniquely if  $T_0[B]$  is known. We have shown it is valid for the hole theory of positrons. But in **I** we have given formulas for calculating  $T_0[B]$  whose correctness relative to the hole theory we have there demonstrated. Hence we have shown that the  $T_{e2}[B]$  obtained by solving (45) with the initial condition  $T_0[B]$  as given by the rules in **I** will be equal to that given for the same problem by the second quantization theory of the Dirac matter field coupled with the quantized electromagnetic field. But it is evident (the argument is given in the next paragraph) that the rules<sup>16</sup> given in **II** constitute a solution in power series in  $e^2$  of the Eq. (45) [which for  $e^2=0$  reduce to the  $T_0[B]$  given in **I**]. Hence the rules in **II** must give, to each order in  $e^2$ , the matrix element for any process that would be calculated by the usual theory of second quantization of the matter and electromagnetic fields. This is what we aimed to prove.

That the rules of **II** represent, in a power series expansion, a solution of (45) is clear. For the rules there given may be stated as follows: Suppose that we have a process to order  $k$  in  $e^2$  (i.e., having  $k$  virtual photons) and order  $n$  in the external potential  $B_{\mu}$ . Then, the matrix element for the process with one more virtual photon and two less potentials is that obtained from

<sup>16</sup> That is, of course, those rules of **II** which apply to the unmodified electrodynamics of Dirac electrons. (The limitation excluding real photons in the initial and final states is removed in Sec. 8.) The same arguments clearly apply to nucleons interacting via neutral vector mesons, vector coupling. Other couplings require a minor extension of the argument. The modification to the  $(x_{i+1}|x_i)_e$ , as in (37), produced by some couplings cannot very easily be written without using operators in the exponents. These operators can be treated as numbers if their order of operation is maintained to be always their order in time. This idea will be discussed and applied more generally in a succeeding paper.

the previous matrix by choosing from the  $n$  potentials a pair, say  $B_{\mu}(1)$  acting at 1 and  $B_{\nu}(2)$  acting at 2, replacing them by  $i e^2 \delta_{\mu\nu} \delta_{+}(s_{12}^2)$ , adding the results for each way of choosing the pair, and dividing by  $k+1$ , the present number of photons. The matrix with no virtual photons ( $k=0$ ) being given to any  $n$  by the rules of **I**, this permits terms to all orders in  $e^2$  to be derived by recursion. It is evident that the rule in italics is that of **II**, and equally evident that it is a word expression of Eq. (45). [The factor  $\frac{1}{2}$  in (45) arises since in integrating over all  $d\tau_1$  and  $d\tau_2$  we count each pair twice. The division by  $k+1$  is required by the rules of **II** for, there, each diagram is to be taken only once, while in the rule given above we say what to do to add one extra virtual photon to  $k$  others. But which one of the  $k+1$  is to be identified at the last photon added is irrelevant. It agrees with (45) of course for it is canceled on differentiating with respect to  $e^2$  the factor  $(e^2)^{k+1}$  for the  $(k+1)$  photons.]

## 8. GENERALIZED FORMULATION OF QUANTUM ELECTRODYNAMICS

The relation implied by (45) between the formal solution for the amplitude for a process in an arbitrary unquantized external potential to that in a quantized field appears to be of much wider generality. We shall discuss the relation from a more general point of view here (still limiting ourselves to the case of no photons in initial or final state).

In earlier sections we pointed out that as a consequence of the Lagrangian form of quantum mechanics the aspects of the particles' motions and the behavior of the field could be analyzed separately. What we did was to integrate over the field oscillator coordinates first. We could, in principle, have integrated over the particle variables first. That is, we first solve the problem with the action of the particles and their interaction with the field and then multiply by the exponential of the action of the field and integrate over all the field oscillator coordinates. (For simplicity of discussion let us put aside from detailed special consideration the questions involving the separation of the longitudinal and transverse parts of the field.<sup>9</sup>) Now the integral over the particle coordinates for a given process is precisely the integral required for the analysis of the motion of the particles in an unquantized potential. With this observation we may suggest a generalization to all types of systems.

Let us suppose the formal solution for the amplitude for some given process with matter in an external potential  $B_{\mu}(1)$  is some numerical quantity  $T_0$ . We mean matter in a more general sense now, for the motion of the matter may be described by the Dirac equation, or by the Klein-Gordon equation, or may involve charged or neutral particles other than electrons and positrons in any manner whatsoever. The quantity  $T_0$  depends of course on the potential function  $B_{\mu}(1)$ ; that is, it is a functional  $T_0[B]$  of this potential. We

assume we have some expression for it in terms of  $B_\mu$  (exact, or to some desired degree of approximation in the strength of the potential).

Then the answer  $T_{\epsilon^2}[B]$  to the corresponding problem in quantum electrodynamics is  $T_0[A_\mu(1) + B_\mu(1)] \times \exp(iS_0)$  summed over all possible distributions of field  $A_\mu(1)$ , wherein  $S_0$  is the action for the field  $S_0 = -(8\pi e^2)^{-1} \sum_\mu \int ((\partial A_\mu/\partial t)^2 - (\nabla A_\mu)^2) d^3x dt$  the sum on  $\mu$  carrying the usual minus sign for space components.

If  $F[A]$  is any functional of  $A_\mu(1)$  we shall represent by  ${}_0|F[A]|_0$  this superposition of  $F[A] \exp(iS_0)$  over distributions of  $A_\mu$  for the case in which there are no photons in initial or final state. That is, we have

$$T_{\epsilon^2}[B] = {}_0|T_0[A+B]|_0. \quad (46)$$

The evaluation of  ${}_0|F[A]|_0$  directly from the definition of the operation  ${}_0| \cdot |_0$  is not necessary. We can give the result in another way. We first note that the operation is linear,

$${}_0|F_1[A] + F_2[A]|_0 = {}_0|F_1[A]|_0 + {}_0|F_2[A]|_0 \quad (47)$$

so that if  $F$  is represented as a sum of terms each term can be analyzed separately. We have studied essentially the case in which  $F[A]$  is an exponential function. In fact, what we have done in Section 4 may be repeated with slight modification to show that

$$\begin{aligned} & {}_0|\exp\left(-i\int j_\mu(1)A_\mu(1)d\tau_1\right)|_0 \\ &= \exp\left(-\frac{1}{2}ie^2\int\int j_\mu(1)j_\nu(2)\delta_+(s_{12}^2)d\tau_1d\tau_2\right) \end{aligned} \quad (48)$$

where  $j_\mu(1)$  is an arbitrary function of position and time for each value of  $\mu$ .

Although this gives the evaluation of  ${}_0| \cdot |_0$  for only a particular functional of  $A_\mu$  the appearance of the arbitrary function  $j_\mu(1)$  makes it sufficiently general to permit the evaluation for any other functional. For it is to be expected that any functional can be represented as a superposition of exponentials with different functions  $j_\mu(1)$  (by analogy with the principle of Fourier integrals for ordinary functions). Then, by (47), the result of the operation is the corresponding superposition of expressions equal to the right-hand side of (48) with the various  $j$ 's substituted for  $j_\mu$ .

In many applications  $F[A]$  can be given as a power series in  $A_\mu$ :

$$\begin{aligned} F[A] &= f_0 + \int f_\mu(1)A_\mu(1)d\tau_1 \\ &+ \int\int f_{\mu\nu}(1, 2)A_\mu(1)A_\nu(2)d\tau_1d\tau_2 + \dots \end{aligned} \quad (49)$$

where  $f_0, f_\mu(1), f_{\mu\nu}(1, 2) \dots$  are known numerical func-

tions independent of  $A_\mu$ . Then by (47)

$$\begin{aligned} {}_0|F[A]|_0 &= f_0 + \int f_\mu(1)_0|A_\mu(1)|_0 d\tau_1 \\ &+ \int\int f_{\mu\nu}(1, 2)_0|A_\mu(1)A_\nu(2)|_0 d\tau_1d\tau_2 + \dots \end{aligned} \quad (50)$$

where we set  ${}_0|1|_0 = 1$  (from (48) with  $j_\mu = 0$ ). We can work out expressions for the successive powers of  $A_\mu$  by differentiating both sides of (48) successively with respect to  $j_\mu$  and setting  $j_\mu = 0$  in each derivative. For example, the first variation (derivative) of (48) with respect to  $j_\mu(3)$  gives

$$\begin{aligned} & {}_0\left|-iA_\mu(3)\exp\left(-i\int j_\nu(1)A_\nu(1)d\tau_1\right)\right|_0 \\ &= -ie^2\int\delta_+(s_{34}^2)j_\mu(4)d\tau_4 \\ &\times \exp\left(-\frac{1}{2}ie^2\int\int j_\nu(1)j_\nu(2)\delta_+(s_{12}^2)d\tau_1d\tau_2\right). \end{aligned} \quad (51)$$

Setting  $j_\mu = 0$  gives

$${}_0|A_\mu(3)|_0 = 0.$$

Differentiating (51) again with respect to  $j_\nu(4)$  and setting  $j_\nu = 0$  shows

$${}_0|A_\mu(3)A_\nu(4)|_0 = ie^2\delta_{\mu\nu}\delta_+(s_{34}^2) \quad (52)$$

and so on for higher powers. These results may be substituted into (50). Clearly therefore when  $T_0[B+A]$  in (46) is expanded in a power series and the successive terms are computed in this way, we obtain the results given in II.

It is evident that (46), (47), (48) imply that  $T_{\epsilon^2}[B]$  satisfies the differential equation (45) and conversely (45) with the definition (46) implies (47) and (48). For if  $T_0[B]$  is an exponential

$$T_0[B] = \exp\left(-i\int j_\mu(1)B_\mu(1)d\tau_1\right) \quad (53)$$

we have from (46), (48) that

$$\begin{aligned} T_{\epsilon^2}[B] &= \exp\left[-\frac{1}{2}ie^2\int\int j_\mu(1)j_\nu(2)\delta_+(s_{12}^2)d\tau_1d\tau_2\right] \\ &\cdot \exp\left[-i\int j_\nu(1)B_\nu(1)d\tau_1\right]. \end{aligned} \quad (54)$$

Direct substitution of this into Eq. (45) shows it to be a solution satisfying the boundary condition (53). Since the differential equation (45) is linear, if  $T_0[B]$  is a superposition of exponentials, the corresponding superposition of solutions (54) is also a solution.

Many of the formal representations of the matter system (such as that of second quantization of Dirac electrons) represent the interaction with a fixed potential in a formal exponential form such as the left-hand side of (48), except that  $j_\mu(1)$  is an operator instead of a numerical function. Equation (48) may still be used if care is exercised in defining the order of the operators on the right-hand side. The succeeding paper will discuss this in more detail.

Equation (45) or its solution (46), (47), (48) constitutes a very general and convenient formulation of the laws of quantum electrodynamics for virtual processes. Its relativistic invariance is evident if it is assumed that the unquantized theory giving  $T_0[B]$  is invariant. It has been proved to be equivalent to the usual formulation for Dirac electrons and positrons (for Klein-Gordon particles see Appendix A). It is suggested that it is of wide generality. It is expressed in a form which has meaning even if it is impossible to express the matter system in Hamiltonian form; in fact, it only requires the existence of an amplitude for fixed potentials which obeys the principle of superposition of amplitudes. If  $T_0[B]$  is known in power series in  $B$ , calculations of  $T_{e2}[B]$  in a power series of  $e^2$  can be made directly using the italicized rule of Sec. 7. The limitation to virtual quanta is removed in the next section.

On the other hand, the formulation is unsatisfactory because for situations of importance it gives divergent results, even if  $T_0[B]$  is finite. The modification proposed in II of replacing  $\delta_+(s_{12}^2)$  in (45), (48) by  $f_+(s_{12}^2)$  is not satisfactory owing to the loss of the theorems of conservation of energy or probability discussed in II at the end of Sec. 6. There is the additional difficulty in positron theory that even  $T_0[B]$  is infinite to begin with (vacuum polarization). Computational ways of avoiding these troubles are given in II and in the references of footnote 2.

#### 9. CASE OF REAL PHOTONS

The case in which there are real photons in the initial or the final state can be worked out from the beginning in the same manner.<sup>17</sup> We first consider the case of a system interacting with a single oscillator. From this result the generalization will be evident. This time we shall calculate the transition element between an initial state in which the particle is in state  $\psi_{t'}$  and the oscillator is in its  $n$ th eigenstate (i.e., there are  $n$  photons in the field) to a final state with particle in  $\chi_{t''}$ , oscillator in  $m$ th level. As we have already discussed, when the coordinates of the oscillator are eliminated the result is the transition element  $\langle \chi_{t''} | G_{mn} | \psi_{t'} \rangle$  where

$$G_{mn} = \int \varphi_m^*(q_j) k(q_j, t''; q_0, t') \varphi_n(q_0) dq_0 dq_j \quad (11)$$

where  $\varphi_m$ ,  $\varphi_n$  are the wave functions<sup>8</sup> for the oscillator

<sup>17</sup> For an alternative method starting directly from the formula (24) for virtual photons, see Appendix B.

in state  $m$ ,  $n$  and  $k$  is given in (12). The  $G_{mn}$  can be evaluated most easily by calculating the generating function

$$g(X, Y) = \sum_m \sum_n G_{mn} X^m Y^n (m! n!)^{-\frac{1}{2}} \quad (55)$$

for arbitrary  $X$ ,  $Y$ . If expression (11) is substituted in the left-hand side of (55), the expression can be simplified by use of the generating function relation for the eigenfunctions<sup>8</sup> of the harmonic oscillator

$$\begin{aligned} \sum_n \varphi_n(q_0) Y^n (n!)^{-\frac{1}{2}} &= (\omega/\pi)^{\frac{1}{2}} \exp(-\frac{1}{2}i\omega t') \\ &\times \exp^{\frac{1}{2}}[\omega q_0^2 - (Y \exp[-i\omega t'] - (2\omega)^{\frac{1}{2}} q_0)^2] \end{aligned}$$

Using a similar expansion for the  $\varphi_m^*$  one is left with the exponential of a quadratic function of  $q_0$  and  $q_j$ . The integration on  $q_0$  and  $q_j$  is then easily performed to give

$$g(X, Y) = G_{00} \exp(XY + i\beta^* X + i\beta Y) \quad (56)$$

from which expansion in powers of  $X$  and  $Y$  and comparison to (11) gives the final result

$$G_{mn} = G_{00} (m! n!)^{-\frac{1}{2}} \sum_r \frac{m!}{r} \frac{n!}{(m-r)! r! (n-r)! r!} \times r! (i\beta^*)^{m-r} (i\beta)^{n-r} \quad (57)$$

where  $G_{00}$  is given in (14) and

$$\begin{aligned} \beta &= (2\omega)^{-\frac{1}{2}} \int_{t'}^{t''} \gamma(t) \exp(-i\omega t) dt, \\ \beta^* &= (2\omega)^{-\frac{1}{2}} \int_{t'}^{t''} \gamma(t) \exp(+i\omega t) dt, \end{aligned} \quad (58)$$

and the sum on  $r$  is to go from 0 to  $m$  or to  $n$  whichever is the smaller. (The sum can be expressed as a Laguerre polynomial but there is no advantage in this.)

Formula (57) is readily understandable. Consider first a simple case of absorption of one photon. Initially we have one photon and finally none. The amplitude for this is the transition element of  $G_{01} = i\beta G_{00}$  or  $\langle \chi_{t'} | i\beta G_{00} | \psi_{t'} \rangle$ . This is the same as would result if we asked for the transition element for a problem in which all photons are virtual but there was present a perturbing potential  $-(2\omega)^{-\frac{1}{2}} \gamma(t) \exp(-i\omega t)$  and we required the first-order effect of this potential. Hence photon absorption is like the first order action of a potential varying in time as  $\gamma(t) \exp(-i\omega t)$  that is with a positive frequency (i.e., the sign of the coefficient of  $t$  in the exponential corresponds to positive energy). The amplitude for emission of one photon involves  $G_{10} = i\beta^* G_{00}$ , which is the same result except that the potential has negative frequency. Thus we begin by interpreting  $i\beta^*$  as the amplitude for emission of one photon  $i\beta$  as the amplitude for absorption of one.

Next for the general case of  $n$  photons initially and  $m$  finally we may understand (57) as follows. We first

neglect Bose statistics and imagine the photons as individual distinct particles. If we start with  $n$  and end with  $m$  this process may occur in several different ways. The particle may absorb in total  $n-r$  of the photons and the final  $m$  photons will represent  $r$  of the photons which were present originally plus  $m-r$  new photons emitted by the particle. In this case the  $n-r$  which are to be absorbed may be chosen from among the original  $n$  in  $n!/(n-r)!r!$  different ways, and each contributes a factor  $i\beta$ , the amplitude for absorption of a photon. Which of the  $m-r$  photons from among the  $m$  are emitted can be chosen in  $m!/(m-r)!r!$  different ways and each photon contributes a factor  $i\beta^*$  in amplitude. The initial  $r$  photons which do not interact with the particle can be re-arranged among the final  $r$  in  $r!$  ways. We must sum over the alternatives corresponding to different values of  $r$ . Thus the form of  $G_{mn}$  can be understood. The remaining factor  $(m!)^{-1}(n!)^{-1}$  may be interpreted as saying that in computing probabilities (which therefore involves the square of  $G_{mn}$ ) the photons may be considered as independent but that if  $m$  are actually equal the statistical weight of each of the states which can be made by rearranging the  $m$  equal photons is only  $1/m!$ . This is the content of Bose statistics; that  $m$  equal particles in a given state represents just one state, i.e., has statistical weight unity, rather than the  $m!$  statistical weight which would result if it is imagined that the particles and states can be identified and rearranged in  $m!$  different ways. This holds for both the initial and final states of course. From this rule about the statistical weights of states the derivation of the blackbody distribution law follows.

The actual electromagnetic field is represented as a host of oscillators each of which behaves independently and produces its own factor such as  $G_{mn}$ . Initial or final states may also be linear combinations of states in which one or another oscillator is excited. The results for this case are of course the corresponding linear combination of transition elements.

For photons of a given direction of polarization and for sin or cos waves the explicit expression for  $\beta$  can be obtained directly from (58) by substituting the formulas (16) for the  $\gamma$ 's for the corresponding oscillator. It is more convenient to use the linear combination corresponding to running waves. Thus we find the amplitude for absorption of a photon of momentum  $\mathbf{K}$ , frequency  $k=(\mathbf{K}\cdot\mathbf{K})^{1/2}$  polarized in direction  $\mathbf{e}$  is given by including a factor  $i$  times

$$\begin{aligned} \beta_{\mathbf{K},\mathbf{e}} = & (4\pi)^{1/2}(2k)^{-1} \sum_n e_n \int_{t'}^{t''} \exp(-ikt) \\ & \times \exp(i\mathbf{K}\cdot\mathbf{x}_n(t)) \mathbf{e}\cdot\mathbf{x}'_n(t) dt \quad (59) \end{aligned}$$

in the transition element (25). The density of states in momentum space is now  $(2\pi)^{-3}d^3\mathbf{K}$ . The amplitude for emission is just  $i$  times the complex conjugate of

this expression, or what amounts to the same thing, the same expression with the sign of the four vector  $k_\mu$  reversed. Since the factor (59) is exactly the first-order effect of a vector potential

$$A^{PH} = (2\pi/k)^{1/2} \mathbf{e} \exp(-i(kt - \mathbf{K}\cdot\mathbf{x}))$$

of the corresponding classical wave, we have derived the rules for handling real photons discussed in II.

We can express this directly in terms of the quantity  $T_{e2}[B]$ , the amplitude for a given transition without emission of a photon. What we have said is that the amplitude for absorption of just one photon whose classical wave form is  $A_\mu^{PH}(1)$  (time variation  $\exp(-ik\tau_1)$  corresponding to positive energy  $k$ ) is proportional to the first order (in  $\epsilon$ ) change produced in  $T_{e2}[B]$  on changing  $B$  to  $B + \epsilon A^{PH}$ . That is, more exactly,

$$\int (\delta T_{e2}[B]/\delta B_\mu(1)) A_\mu^{PH}(1) d\tau_1 \quad (60)$$

is the amplitude for absorption by the particle system of one photon,  $A^{PH}$ . (A superposition argument shows the expression to be valid not only for plane waves, but for spherical waves, etc., as given by the form of  $A^{PH}$ .) The amplitude for emission is the same expression but with the sign of the frequency reversed in  $A^{PH}$ . The amplitude that the system absorbs two photons with waves  $A_\mu^{PH_1}$  and  $A_\nu^{PH_2}$  is obtained from the next derivative,

$$\int \int (\delta^2 T_{e2}[B]/\delta B_\mu(1) \delta B_\nu(2)) A_\mu^{PH_1}(1) A_\nu^{PH_2}(2) d\tau_1 d\tau_2,$$

the same expression holding for the absorption of one and emission of the other, or emission of both depending on the sign of the time dependence of  $A^{PH_1}$  and  $A^{PH_2}$ . Larger photon numbers correspond to higher derivatives, absorption of  $l_1$  emission of  $l_2$  requiring the  $(l_1+l_2)$  derivatives. When two or more of the photons are exactly the same (e.g.,  $A^{PH_1}=A^{PH_2}$ ) the same expression holds for the amplitude that  $l_1$  are absorbed by the system while  $l_2$  are emitted. However, the statement that initially  $n$  of a kind are present and  $m$  of this kind are present finally, does not imply  $l_1=n$  and  $l_2=m$ . It is possible that only  $n-r=l_1$  were absorbed by the system and  $m-r=l_2$  emitted, and that  $r$  remained from initial to final state without interaction. This term is weighed by the combinatorial coefficient  $(m!n!)^{-1} \binom{m}{r} \binom{n}{r} r!$  and summed over the possibilities for  $r$  as explained in connection with (57). Thus once the amplitude for virtual processes is known, that for real photon processes can be obtained by differentiation.

It is possible, of course, to deal with situations in which the electromagnetic field is not in a definite state after the interaction. For example, we might ask for the total probability of a given process, such as a scattering, without regard for the number of photons emitted. This is done of course by squaring the ampli-

tude for the emission of  $m$  photons of a given kind and summing on all  $m$ . Actually the sums and integrations over the oscillator momenta can usually easily be performed analytically. For example, the amplitude, starting from vacuum and ending with  $m$  photons of a given kind, is by (56) just

$$G_{m0} = (m!)^{-\frac{1}{2}} G_{00} (i\beta^*)^m. \quad (61)$$

The square of the amplitude summed on  $m$  requires the product of two such expressions (the  $\gamma(t)$  in the  $\beta$  of one and in the other will have to be kept separately) summed on  $m$ :

$$\sum_m G_{m0}^* G_{m0}' = \sum_m G_{00}^* G_{00}' (m!)^{-\frac{1}{2}} \beta^m (\beta'^*)^m = G_{00}^* G_{00}' \exp(\beta\beta'^*).$$

In the resulting expression the sum over all oscillators is easily done. Such expressions can be of use in the analysis in a direct manner of problems of line width, of the Bloch-Nordsieck infra-red problem, and of statistical mechanical problems, but no such applications will be made here.

The author appreciates his opportunities to discuss these matters with Professor H. A. Bethe and Professor J. Ashkin, and the help of Mr. M. Baranger with the manuscript.

#### APPENDIX A. THE KLEIN-GORDON EQUATION

In this Appendix we describe a formulation of the equations for a particle of spin zero which was first used to obtain the rules given in II for such particles. The complete physical significance of the equations has not been analyzed thoroughly so that it may be preferable to derive the rules directly from the second quantization formulation of Pauli and Weisskopf. This can be done in a manner analogous to the derivation of the rules for the Dirac equation given in I or from the Schwinger-Tomonaga formulation<sup>2</sup> in a manner described, for example, by Rohrlich.<sup>18</sup> The formulation given here is therefore not necessary for a description of spin zero particles but is given only for its own interest as an alternative to the formulation of second quantization.

We start with the Klein-Gordon equation

$$(i\partial/\partial x_\mu - A_\mu)^2 \psi = m^2 \psi \quad (1A)$$

for the wave function  $\psi$  of a particle of mass  $m$  in a given external potential  $A_\mu$ . We shall try to represent this in a manner analogous to the formulation of quantum mechanics in C. That is, we try to represent the amplitude for a particle to get from one point to another as a sum over all trajectories of an amplitude  $\exp(iS)$  where  $S$  is the classical action for a given trajectory. To maintain the relativistic invariance in evidence the idea suggests itself of describing a trajectory in space-time by giving the four variables  $x_\mu(u)$  as functions of some fifth parameter  $u$  (rather than expressing  $x_1, x_2, x_3$  in terms of  $x_4$ ). As we expect to represent paths which may reverse themselves in time (to represent pair production, etc., as in I) this is certainly a more convenient representation, for all four functions  $x_\mu(u)$  may be considered as functions of a parameter  $u$  (somewhat analogous to proper time) which increase as we go along the trajectory, whether the trajectory is proceeding forward ( $dx_4/du > 0$ ) or backward ( $dx_4/du < 0$ ) in time.<sup>19</sup> We shall

<sup>18</sup> F. Rohrlich (to be published).

<sup>19</sup> The physical ideas involved in such a description are discussed in detail by Y. Nambu, Prog. Theor. Phys. 5, 82 (1950). An equation of type (2A) extended to the case of Dirac electrons has been studied by V. Fock, Physik Zeits. Sowjetunion 12, 404 (1937).

then have a new type of wave function  $\varphi(x, u)$  a function of five variables,  $x$  standing for the four  $x_\mu$ . It gives the amplitude for arrival at point  $x_\mu$  with a certain value of the parameter  $u$ . We shall suppose that this wave function satisfies the equation

$$i\partial\varphi/\partial u = -\frac{1}{2}(i\partial/\partial x_\mu - A_\mu)^2 \varphi \quad (2A)$$

which is seen to be analogous to the time-dependent Schrödinger equation,  $u$  replacing the time and the four coordinates of space-time  $x_\mu$  replacing the usual three coordinates of space.

Since the potentials  $A_\mu(x)$  are functions only of coordinates  $x_\mu$  and are independent of  $u$ , the equation is separable in  $u$  and we can write a special solution in the form  $\varphi = \exp(-\frac{1}{2}im^2u)\psi(x)$  where  $\psi(x)$ , a function of the coordinates  $x_\mu$  only, satisfies (1A) and the eigenvalue  $\frac{1}{2}m^2$  conjugate to  $u$  is related to the mass  $m$  of the particle. Equation (2A) is therefore equivalent to the Klein-Gordon Eq. (1A) provided we ask in the end only for the solution of (1A) corresponding to the eigenvalue  $\frac{1}{2}m^2$  for the quantity conjugate to  $u$ .

We may now proceed to represent Eq. (2A) in Lagrangian form in general and without regard to this eigenvalue condition. Only in the final solutions need we apply the eigenvalue condition. That is, if we have some special solution  $\varphi(x, u)$  of (2A) we can select that part corresponding to the eigenvalue  $\frac{1}{2}m^2$  by calculating

$$\psi(x) = \int_{-\infty}^{\infty} \exp(-\frac{1}{2}im^2u) \varphi(x, u) du$$

and thereby obtain a solution  $\psi$  of Eq. (1A).

Since (2A) is so closely analogous to the Schrödinger equation, it is easily written in the Lagrangian form described in C, simply by working by analogy. For example if  $\varphi(x, u)$  is known at one value of  $u$  its value at a slightly larger value  $u+\epsilon$  is given by

$$\varphi(x, u+\epsilon) = \int \exp[i\epsilon \left( -\frac{(x_\mu - x_\mu')^2}{2\epsilon^2} - \frac{1}{2} \frac{(x_\mu - x_\mu')}{\epsilon} (A_\mu(x) + A_\mu(x')) \right)] \cdot \varphi(x', u) d^4\tau_{x'} (2\pi i\epsilon)^{-\frac{1}{2}} (-2\pi i\epsilon)^{-\frac{1}{2}} \quad (3A)$$

where  $(x_\mu - x_\mu')^2$  means  $(x_\mu - x_\mu')(x_\mu - x_\mu')$ ,  $d^4\tau_{x'} = dx'_1 dx'_2 dx'_3 dx'_4$  and the sign of the normalizing factor is changed for the  $x_4$  component since the component has the reversed sign in its quadratic coefficient in the exponential, in accordance with our summation convention  $a_\mu b_\mu = a_1 b_4 - a_2 b_1 - a_3 b_2 - a_4 b_3$ . Equation (3A), as can be verified readily as described in C, Sec. 6, is equivalent to first order in  $\epsilon$ , to Eq. (2A). Hence, by repeated use of this equation the wave function at  $u_0 = n\epsilon$  can be represented in terms of that at  $u=0$  by:

$$\begin{aligned} \varphi(x_{n\epsilon}, u_0) = & \int \exp \left[ -\frac{i\epsilon}{2} \sum_{i=1}^n \left[ \left( \frac{x_{\mu,i} - x_{\mu,i-1}}{\epsilon} \right)^2 \right. \right. \\ & \left. \left. + \epsilon^{-1} (x_{\mu,i} - x_{\mu,i-1}) (A_\mu(x_i) + A_\mu(x_{i-1})) \right] \right] \\ & \cdot \varphi(x_{n\epsilon}, 0) \prod_{i=0}^{n-1} (d^4\tau_i / 4\pi^2 \epsilon^2 i). \end{aligned} \quad (4A)$$

That is, roughly, the amplitude for getting from one point to another with a given value of  $u_0$  is the sum over all trajectories of  $\exp(iS)$  where

$$S = - \int_0^{u_0} [\frac{1}{2} (dx_\mu/du)^2 + (dx_\mu/du) A_\mu(x)] du, \quad (5A)$$

when sufficient care is taken to define the quantities, as in C. This completes the formulation for particles in a fixed potential but a few words of description may be in order.

In the first place in the special case of a free particle we can define a kernel  $k^{(0)}(x, u_0; x', 0)$  for arrival from  $x_\mu', 0$  to  $x_\mu$  at  $u_0$  as the sum over all trajectories between these points of  $\exp -i \int_0^{u_0} \frac{1}{2} (dx_\mu/du)^2 du$ . Then for this case we have

$$\varphi(x, u_0) = \int k^{(0)}(x, u_0; x', 0) \varphi(x', 0) d^4\tau_{x'}, \quad (6A)$$

and it is easily verified that  $k_0$  is given by

$$k^{(0)}(x, u_0; x', 0) = (4\pi^2 u_0^2 i)^{-\frac{1}{2}} \exp -i(x_\mu - x_\mu')^2 / 2u_0 \quad (7A)$$

for  $u_0 > 0$  and by 0, by definition, for  $u_0 < 0$ . The corresponding

kernel of importance when we select the eigenvalue  $\frac{1}{2}m^2$  is<sup>20</sup>

$$\begin{aligned} 2I_+(x, x') &= \int_{-\infty}^{\infty} k^{(0)}(x, u_0; x', 0) \exp(-\frac{1}{2}im^2u_0) du_0 \\ &= \int_0^{\infty} du_0 (4\pi^2 u_0^2 i)^{-1} \exp(-\frac{1}{2}i(m^2 u_0 + u_0^{-1}(x_\mu - x'_\mu)^2)) \end{aligned} \quad (8A)$$

(the last extends only from  $u_0=0$  since  $k_0$  is zero for negative  $u_0$ ) which is identical to the  $I_+$  defined in II.<sup>21</sup> This may be seen readily by studying the Fourier transform, for the transform of the integrand on the right-hand side is

$$\begin{aligned} \int (4\pi^2 u_0^2 i)^{-1} \exp(ip \cdot x) \exp(-\frac{1}{2}i(m^2 u_0 + x_\mu^2/u_0)) d^4\tau_x \\ = \exp(-\frac{1}{2}iu_0(m^2 - p_\mu^2)) \end{aligned}$$

so that the  $u_0$  integration gives for the transform of  $I_+$  just  $1/(p_\mu^2 - m^2)$  with the pole defined exactly as in II. Thus we are automatically representing the positrons as trajectories with the time sense reversed.

If  $\Phi^{(0)}[x(u)] = \exp(-i \int_0^{u_0} (dx_\mu/du)^2 du)$  is the amplitude for a given trajectory  $x_\nu(u)$  for a free particle, then the amplitude in a potential is

$$\Phi^{(A)}[x(u)] = \Phi^{(0)}[x(u)] \exp(-i \int_0^{u_0} (dx_\mu/du) A_\mu(x) du). \quad (9A)$$

If desired this may be studied by perturbation methods by expanding the exponential in powers of  $A_\mu$ .

For interpretation, the integral in (9A) must be written as a Riemann sum, and if a perturbation expansion is made, care must be taken with the terms quadratic in the velocity, for the effect of  $(x_{\mu,i+1} - x_{\mu,i})(x_{\nu,i+1} - x_{\nu,i})$  is not of order  $\epsilon^2$  but is  $-i\delta_{\mu\nu}\epsilon$ . The "velocity"  $dx_\mu/du$  becomes the momentum operator  $p_\mu = +i\partial/\partial x_\mu$  operating half before and half after  $A_\mu$ , just as in the non-relativistic Schrödinger equation discussed in Sec. 5. Furthermore, in exactly the same manner as in that case, but here in four dimensions, a term quadratic in  $A_\mu$  arises in the second-order perturbation terms from the coincidence of two velocities for the same value of  $u$ .

As an example, the kernel  $k^{(A)}(x, u_0; x', 0)$  for proceeding from  $x'_\mu$ , 0 to  $x_\mu, u_0$  in a potential  $A_\mu$  differs from  $k^{(0)}$  to first order in  $A_\mu$  by a term

$$-i \int_0^{u_0} du k^{(0)}(x, u_0; y, u) \frac{1}{2} (p_\mu A_\mu(y) + A_\mu(y) p_\mu) k^{(0)}(y, u; x', 0) d\tau_y$$

the  $p_\mu$  here meaning  $+i\partial/\partial y_\mu$ . The kernel of importance on selecting the eigenvalue  $\frac{1}{2}m^2$  is obtained by multiplying this by  $\exp(-\frac{1}{2}im^2u_0)$  and integrating  $u_0$  from 0 to  $\infty$ . The kernel  $k^{(0)}(x, u_0; y, u)$  depends only on  $u' = u_0 - u$  and in the integrals on  $u$  and  $u_0$ ;  $\int_0^{\infty} du_0 \int_0^{u_0} du \exp(-\frac{1}{2}im^2u_0) \dots$ , can be written, on interchanging the order of integration and changing variables to  $u$  and  $u'$ ,  $\int_0^{\infty} du \int_0^{\infty} du' \exp(-\frac{1}{2}im^2(u+u')) \dots$ . Now the integral on  $u'$  converts  $k^{(0)}(x, u_0; y, u)$  to  $2iI_+(x, y)$  by (8A), while that on  $u$  converts  $k^{(0)}(y, u; x', 0)$  to  $2iI_+(y, x')$ , so the result becomes

$$\int 2iI_+(x, y) (p_\mu A_\mu + A_\mu p_\mu) I_+(y, x') d^4\tau_y$$

as expected. The same principle works to any order so that the rules for a single Klein-Gordon particle in external potentials given in II, Section 9, are deduced.

The transition to quantum electrodynamics is simple for in (5A) we already have a transition amplitude represented as a sum (over trajectories, and eventually  $u_0$ ) of terms, in each of which the potential appears in exponential form. We may make use of the general relation (54). Hence, for example, one finds

<sup>20</sup> The factor  $2i$  in front of  $I_+$  is simply to make the definition of  $I_+$  here agree with that in I and II. In II it operates with  $\mathbf{p} \cdot \mathbf{A} + \mathbf{A} \cdot \mathbf{p}$  as a perturbation. But the perturbation coming from (3A) in a natural way by expansion of the exponential is  $-\frac{1}{2}i(\mathbf{p} \cdot \mathbf{A} + \mathbf{A} \cdot \mathbf{p})$ .

<sup>21</sup> Expression (8A) is closely related to Schwinger's parametric integral representation of these functions. For example, (8A) becomes formula (45) of F. Dyson, Phys. Rev. **75**, 486 (1949) for  $\Delta_{\mathbf{p}} \equiv \Delta^{(1)} - 2i\bar{\Delta} \equiv 2iI_+$  if  $(2\alpha)^{-1}$  is substituted for  $u_0$ .

for the case of no photons in the initial and final states, in the presence of an external potential  $B_\mu$ , the amplitude that a particle proceeds from  $(x_\mu, 0)$  to  $(x_\mu, u_0)$  is the sum over all trajectories of the quantity

$$\exp -i \left[ \frac{1}{2} \int_0^{u_0} \left( \frac{dx_\mu}{du} \right)^2 du + \int_0^{u_0} \frac{dx_\mu}{du} B_\mu(x(u)) du \right. \\ \left. + \frac{e^2}{2} \int_0^{u_0} \int_0^{u_0} \frac{dx_\mu(u)}{du} \frac{dx_\nu(u')}{du'} \delta_+((x_\mu(u) - x_\mu(u'))^2) du du' \right]. \quad (10A)$$

This result must be multiplied by  $\exp(-\frac{1}{2}im^2u_0)$  and integrated on  $u_0$  from zero to infinity to express the action of a Klein-Gordon particle acting on itself through virtual photons. The integrals are interpreted as Riemann sums, and if perturbation expansions are made, the necessary care is taken with the terms quadratic in velocity. When there are several particles (other than the virtual pairs already included) one uses a separate  $u$  for each, and writes the amplitude for each set of trajectories as the exponential of  $-i$  times

$$\begin{aligned} \frac{1}{2} \sum_n \int_0^{u_0^{(n)}} \left( \frac{dx_\mu^{(n)}}{du} \right)^2 du + \sum_n \int_0^{u_0^{(n)}} \frac{dx_\mu^{(n)}}{du} B_\mu(x_\mu^{(n)}(u)) du \\ + \frac{e^2}{2} \sum_{nm} \sum_0^{u_0^{(n)}} \int_0^{u_0^{(n)}} \frac{dx_\mu^{(n)}(u)}{du} \frac{dx_\nu^{(m)}(u')}{du'} \\ \times \delta_+((x_\mu^{(n)}(u) - x_\mu^{(m)}(u'))^2) du du', \end{aligned} \quad (11A)$$

where  $x_\mu^{(n)}(u)$  are the coordinates of the trajectory of the  $n$ th particle.<sup>22</sup> The solution should depend on the  $u_0^{(n)}$  as  $\exp(-\frac{1}{2}im^2 \sum_n u_0^{(n)})$ .

Actually, knowledge of the motion of a single charge implies a great deal about the behavior of several charges. For a pair which eventually may turn out to be a virtual pair may appear in the short run as two "other particles." As a virtual pair, that is, as the reverse section of a very long and complicated single track we know its behavior by (10A). We can assume that such a section can be looked at equally well, for a limited duration at least, as being due to other unconnected particles. This then implies a definite law of interaction of particles if the self-action (10A) of a single particle is known. (This is similar to the relation of real and virtual photon processes discussed in detail in Appendix B.) It is possible that a detailed analysis of this could show that (10A) implied that (11A) was correct for many particles. There is even reason to believe that the law of Bose-Einstein statistics and the expression for contributions from closed loops could be deduced by following this argument. This has not yet been analyzed completely, however, so we must leave this formulation in an incomplete form. The expression for closed loops should come out to be  $C_v = \exp + L$  where  $L$ , the contribution from a

<sup>22</sup> The form (10A) suggests another interesting possibility for avoiding the divergences of quantum electrodynamics in this case. The divergences arise from the  $\delta_+$  function when  $u=u'$ . We might restrict the integration in the double integral such that  $|u-u'| > \delta$  where  $\delta$  is some finite quantity, very small compared with  $m^{-2}$ . More generally, we could keep the region  $u=u'$  from contributing by including in the integrand a factor  $F(u-u')$  where  $F(x) \rightarrow 1$  for  $x$  large compared to some  $\delta$ , and  $F(0)=0$  (e.g.,  $F(x)$  acts qualitatively like  $1 - \exp(-x^2\delta^{-2})$ ). (Another way might be to replace  $u$  by a discontinuous variable, that is, we do not use the limit in (4A) as  $\epsilon \rightarrow 0$  but set  $\epsilon=\delta$ .) The idea is that two interactions would contribute very little in amplitude if they followed one another too rapidly in  $u$ . It is easily verified that this makes the otherwise divergent integrals finite. But whether the resulting formulas make good physical sense is hard to see. The action of a potential would now depend on the value of  $u$  so that Eq. (2A), or its equivalent, would not be separable in  $u$  so that  $\frac{1}{2}m^2$  would no longer be a strict eigenvalue for all disturbances. High energy potentials could excite states corresponding to other eigenvalues, possibly thereby corresponding to other masses. This note is meant only as a speculation, for not enough work has been done in this direction to make sure that a reasonable physical theory can be developed along these lines. (What little work has been done was not promising.) Analogous modifications can also be made for Dirac electrons.

single loop, is

$$L = 2 \int_0^\infty l(u_0) \exp(-\frac{1}{2}im^2 u_0) du_0 / u_0$$

where  $l(u_0)$  is the sum over all trajectories which close on themselves ( $x_\mu(u_0) = x_\mu(0)$ ) of  $\exp(iS)$  with  $S$  given in (5A), and a final integration  $d\tau_{x(0)}$  on  $x_\mu(0)$  is made. This is equivalent to putting

$$l(u_0) = \int (k^{(A)}(x, u_0; x, 0) - k^{(0)}(x, u_0; x, 0)) d\tau_x.$$

The term  $k^{(0)}$  is subtracted only to simplify convergence problems (as adding a constant independent of  $A_\mu$  to  $L$  has no effect).

## APPENDIX B. THE RELATION OF REAL AND VIRTUAL PROCESSES

If one has a general formula for all virtual processes he should be able to find the formulas and states involved in real processes. That is to say, we should be able to deduce the formulas of Section 9 directly from the formulation (24), (25) (or its generalized equivalent such as (46), (48)) without having to go all the way back to the more usual formulation. We discuss this problem here.

That this possibility exists can be seen from the consideration that what looks like a real process from one point of view may appear as a virtual process occurring over a more extended time.

For example, if we wish to study a given real process, such as the scattering of light, we can, if we wish, include in principle the source, scatterer, and eventual absorber of the scattered light in our analysis. We may imagine that no photon is present initially, and that the source then emits light (the energy coming say from kinetic energy in the source). The light is then scattered and eventually absorbed (becoming kinetic energy in the absorber). From this point of view the process is virtual; that is, we start with no photons and end with none. Thus we can analyze the process by means of our formula for virtual processes, and obtain the formulas for real processes by attempting to break the analysis into parts corresponding to emission, scattering, and absorption.<sup>23</sup>

To put the problem in a more general way, consider the amplitude for some transition from a state empty of photons far in the past (time  $t'$ ) to a similar one far in the future ( $t=t''$ ). Suppose the time interval to be split into three regions  $a$ ,  $b$ ,  $c$  in some convenient manner, so that region  $b$  is an interval  $t_2 > t > t_1$  around the present time that we wish to study. Region  $a$ , ( $t_1 > t > t'$ ), precedes  $b$ , and  $c$ , ( $t'' > t > t_2$ ), follows  $b$ . We want to see how it comes about that the phenomena during  $b$  can be analyzed by a study of transitions  $g_{ji}(b)$  between some initial state  $i$  at time  $t_1$  (which no longer need be photon-free), to some other final state  $j$  at time  $t_2$ . The states  $i$  and  $j$  are members of a large class which we will have to find out how to specify. (The single index  $i$  is used to represent a large number of quantum numbers, so that different values of  $i$  will correspond to having various numbers of various kinds of photons in the field, etc.) Our problem is to represent the over-all transition amplitude,  $g(a, b, c)$ , as a sum over various values of  $i, j$  of a product of three amplitudes,

$$g(a, b, c) = \sum_i \sum_j g_{ji}(c) g_{ji}(b) g_{ji}(a); \quad (1B)$$

first the amplitude that during the interval  $a$  the vacuum state makes transition to some state  $i$ , then the amplitude that during  $b$  the transition to  $j$  is made, and finally in  $c$  the amplitude that the transition from  $j$  to some photon-free state 0 is completed.

<sup>23</sup> The formulas for real processes deduced in this way are strictly limited to the case in which the light comes from sources which are originally dark, and that eventually all light emitted is absorbed again. We can only extend it to the case for which these restrictions do not hold by hypothesis, namely, that the details of the scattering process are independent of these characteristics of the light source and of the eventual disposition of the scattered light. The argument of the text gives a method for discovering formulas for real processes when no more than the formula for virtual processes is at hand. But with this method belief in the general validity of the resulting formulas must rest on the physical reasonableness of the above-mentioned hypothesis.

The mathematical problem of splitting  $g(a, b, c)$  is made definite by the further condition that  $g_{ji}(b)$  for given  $i, j$  must not involve the coordinates of the particles for times corresponding to regions  $a$  or  $c$ ,  $g_{i0}(a)$  must involve those only in region  $a$ , and  $g_{j0}(c)$  only in  $c$ .

To become acquainted with what is involved, suppose first that we do not have a problem involving virtual photons, but just the transition of a one-dimensional Schrödinger particle going in a long time interval from, say, the origin  $o$  to the origin  $o$ , and ask what states  $i$  we shall need for intermediary time intervals. We must solve the problem (1B) where  $g(a, b, c)$  is the sum over all trajectories going from  $o$  at  $t'$  to  $o$  at  $t''$  of  $\exp(iS)$  where  $S = \int L dt$ . The integral may be split into three parts  $S = S_a + S_b + S_c$  corresponding to the three ranges of time. Then  $\exp(iS) = \exp(iS_a) \cdot \exp(iS_b) \cdot \exp(iS_c)$  and the separation (1B) is accomplished by taking for  $g_{i0}(a)$  the sum over all trajectories lying in  $a$  from  $o$  to some end point  $x_{t_1}$  of  $\exp(iS_a)$ , for  $g_{ji}(b)$  the sum over trajectories in  $b$  of  $\exp(iS_b)$  between end points  $x_{t_1}$  and  $x_{t_2}$ , and for  $g_{j0}(c)$  the sum of  $\exp(iS_c)$  over the section of the trajectory lying in  $c$  and going from  $x_{t_2}$  to  $o$ . Then the sum on  $i$  and  $j$  can be taken to be the integrals on  $x_{t_1}, x_{t_2}$  respectively. Hence the various states  $i$  can be taken to correspond to particles being at various coordinates  $x$ . (Of course any other representation of the states in the sense of Dirac's transformation theory could be used equally well. Which one, whether coordinate, momentum, or energy level representation, is of course just a matter of convenience and we cannot determine that simply from (1B).)

We can consider next the problem including virtual photons. That is,  $g(a, b, c)$  now contains an additional factor  $\exp(iR)$  where  $R$  involves a double integral  $\int \int$  over all time. Those parts of the index  $i$  which correspond to the particle states can be taken in the same way as though  $R$  were absent. We study now the extra complexities in the states produced by splitting the  $R$ . Let us first (solely for simplicity of the argument) take the case that there are only two regions  $a, c$  separated by time  $t_0$  and try to expand

$$g(a, c) = \sum_i g_{i0}(c) g_{i0}(a).$$

The factor  $\exp(iR)$  involves  $R$  as a double integral which can be split into three parts  $\int_a \int_a + \int_c \int_c + \int_a \int_c$  for the first of which both  $t, s$  are in  $a$ , for the second both are in  $c$ , for the third one is in  $a$  the other in  $c$ . Writing  $\exp(iR)$  as  $\exp(iR_{cc}) \cdot \exp(iR_{aa})$

$\exp(iR_{ac})$  shows that the factors  $R_{cc}$  and  $R_{aa}$  produce no new problems for they can be taken bodily into  $g_{i0}(c)$  and  $g_{i0}(a)$  respectively. However, we must disentangle the variables which are mixed up in  $\exp(iR_{ac})$ .

The expression for  $R_{ac}$  is just twice (24) but with the integral on  $s$  extending over the range  $a$  and that for  $t$  extending over  $c$ . Thus  $\exp(iR_{ac})$  contains the variables for times in  $a$  and in  $c$  in a quite complicated mixture. Our problem is to write  $\exp(iR_{ac})$  as a sum over possibly a vast class of states  $i$  of the product of two parts, like  $h_i(c) h_i(a)$ , each of which involves the coordinates in one interval alone.

This separation may be made in many different ways, corresponding to various possible representations of the state of the electromagnetic field. We choose a particular one. First we can expand the exponential,  $\exp(iR_{ac})$ , in a power series, as  $\sum_n i^n (n!)^{-1} (R_{ac})^n$ . The states  $i$  can therefore be subdivided into subclasses corresponding to an integer  $n$  which we can interpret as the number of quanta in the field at time  $t_0$ . The amplitude for the case  $n=0$  clearly just involves  $\exp(iR_{aa})$  and  $\exp(iR_{cc})$  in the way that it should if we interpret these as the amplitudes for regions  $a$  and  $c$ , respectively, of making a transition between a state of zero photons and another state of zero photons.

Next consider the case  $n=1$ . This implies an additional factor in the transitional element; the factor  $R_{ac}$ . The variables are still mixed up. But an easy way to perform the separation suggests itself. Namely, expand the  $\delta_+((t-s)^2 - (x_n(t) - x_m(s))^2)$  in  $R_{ac}$  as a Fourier integral as

$$\int \exp(-ik|t-s|) \exp(-i\mathbf{K} \cdot (\mathbf{x}_n(t) - \mathbf{x}_m(s))) d^3\mathbf{K} / 4\pi^2 k.$$

For the exponential can be written immediately as a product of  $\exp + i(\mathbf{K} \cdot \mathbf{x}_m(s))$ , a function only of coordinates for times  $s$  in  $a$  (suppose  $s < t$ ), and  $\exp - i\mathbf{K} \cdot \mathbf{x}_n(t)$  (a function only of coordinates during interval  $c$ ). The integral on  $d^3\mathbf{K}$  can be symbolized as a sum over states  $i$  characterized by the value of  $\mathbf{K}$ . Thus the state with  $n=1$  must be further characterized by specifying a vector  $\mathbf{K}$ , interpreted as the momentum of the photon. Finally the factor  $(1 - \mathbf{x}'_n(t) \cdot \mathbf{x}'_m(s))$  in  $R_{ac}$  is simply the sum of four parts each of which is already split (namely 1, and each of the three components in the vector scalar product). Hence each photon of momentum  $\mathbf{K}$  must still be characterized by specifying it as one of four varieties; that is, there are four polarizations.<sup>24</sup> Thus in trying to represent the effect of the past  $a$  on the future  $c$  we are lead to invent photons of four polarizations and characterized by a propagation vector  $\mathbf{K}$ .

The term for a given polarization and value of  $\mathbf{K}$  (for  $n=1$ ) is clearly just  $-\beta_a \beta_a^*$  where the  $\beta_a$  is defined in (59) but with the time integral extending just over region  $a$ , while  $\beta_c$  is the same expression with the integration over region  $c$ . Hence the amplitude for transition during interval  $a$  from a state with no quanta to a state with one in a given state of polarization and momentum is calculated by inclusion of an extra factor  $i\beta_a^*$  in the transition element. Absorption in region  $c$  corresponds to a factor  $i\beta_c$ .

We next turn to the case  $n=2$ . This requires analysis of  $R_{ac}$ .<sup>2</sup> The  $\delta_+$  can be expanded again as a Fourier integral, but for each of the two  $\delta_+$  in  $\frac{1}{2}R_{ac}$  we have a value of  $\mathbf{K}$  which may be different. Thus we say, we have two photons, one of momentum  $\mathbf{K}$  and one momentum  $\mathbf{K}'$  and we sum over all values of  $\mathbf{K}$  and  $\mathbf{K}'$ . (Similarly each photon is characterized by its own independent polarization index.) The factor  $\frac{1}{2}$  can be taken into account neatly by asserting that we count each possible pair of photons as constituting just one state at time  $t_0$ . Then the  $\frac{1}{2}$  arises for the sum over all  $\mathbf{K}, \mathbf{K}'$  (and polarizations) counts each pair twice. On the other hand, for the terms representing two identical photons ( $\mathbf{K}=\mathbf{K}'$ ) of like polarization, the  $\frac{1}{2}$  cannot be so interpreted. Instead we invent the rule that a state of two like photons has statistical weight  $\frac{1}{2}$  as great as that calculated as though the photons were different. This, generalized to  $n$  identical photons, is the rule of Bose statistics.

The higher values of  $n$  offer no problem. The  $1/n!$  is interpreted combinatorially for different photons, and as a statistical factor when some are identical. For example, for all  $n$  identical one obtains a factor  $(n!)^{-1}(-\beta_a \beta_a^*)^n$  so that  $(n!)^{-1}(i\beta_a^*)^n$  can be interpreted as the amplitude for emission (from no initial photons) of  $n$  identical photons, in complete agreement with (61) for  $C_{n0}$ .

To obtain the amplitude for transitions in which neither the initial nor the final state is empty of photons we must consider the more general case of the division into three time regions (IB). This time we see that the factor which involves the coordinates in an entangled manner is  $\exp(i(R_{ab}+R_{bc}+R_{ac}))$ . It is to be expanded in the form  $\Sigma_i \Sigma_j h_i''(c) h_j'(b) h_j(a)$ . Again the expansion in power series and development in Fourier series with a polarization sum will solve the problem. Thus the exponential is  $\Sigma_r \Sigma_{l_1} \Sigma_{l_2} (iR_{ac})^{l_1} (iR_{bc})^{l_2} (l_1!)^{-1} (l_2!)^{-1} (r!)^{-1}$ . Now the  $R$  are written as Fourier series, one of the terms containing  $l_1+l_2+r$  variables  $\mathbf{K}$ . Since  $l_1+r$  involve  $a$ ,  $l_2+r$  involve  $c$  and  $l_1+l_2$  involve  $b$ , this term will give the amplitude that  $l_1+r$  photons are emitted during the interval  $a$ , of those  $l_1$  are absorbed during  $b$  but the remaining  $r$ , along with  $l_2$  new ones emitted during  $b$  go on to be absorbed during the interval  $c$ . We have therefore  $n=l_1+r$  photons in the state at time  $t_1$  when  $b$  begins, and  $m=l_2+r$  at  $t_2$  when  $b$  is over. They each are characterized by momentum vectors and polarizations. When these are different the factors  $(l_1!)^{-1} (l_2!)^{-1} (r!)^{-1}$  are absorbed combinatorially. When some are equal we must invoke the rule of the statistical weights. For

<sup>24</sup> Usually only two polarizations transverse to the propagation vector  $\mathbf{K}$  are used. This can be accomplished by a further rearrangement of terms corresponding to the reverse of the steps leading from (17) to (19). We omit the details here as it is well-known that either formulation gives the same results. See II, Section 8.

example, suppose all  $l_1+l_2+r$  photons are identical. Then  $R_{ab}=i\beta_b \beta_a^*$ ,  $R_{bc}=i\beta_c \beta_b^*$ ,  $R_{ac}=i\beta_a \beta_a^*$  so that our sum is

$$\Sigma_{l_1} \Sigma_{l_2} \Sigma_r (l_1! l_2! r!)^{-1} (i\beta_c)^{l_2+r} (i\beta_b)^{l_1} (i\beta_a^*)^{l_2} (i\beta_a^*)^{l_1+r}.$$

Putting  $m=l_2+r$ ,  $n=l_1+r$ , this is the sum on  $n$  and  $m$  of

$$(i\beta_c)^m (m!)^{-1} [\Sigma_r (m! n!)^{\frac{1}{2}} ((m-r)! (n-r)! r!)^{-1} \times (i\beta_b^*)^{m-r} (i\beta_b)^{n-r}] (n!)^{-1} (i\beta_a^*)^n.$$

The last factor we have seen is the amplitude for emission of  $n$  photons during interval  $a$ , while the first factor is the amplitude for absorption of  $m$  during  $c$ . The sum is therefore the factor for transition from  $n$  to  $m$  identical photons, in accordance with (57). We see the significance of the simple generating function (56).

We have therefore found rules for real photons in terms of those for virtual. The real photons are a way of representing and keeping track of those aspects of the past behavior which may influence the future.

If one starts from a theory involving an arbitrary modification of the direct interaction  $\delta_+$  (or in more general situations) it is possible in this way to discover what kinds of states and physical entities will be involved if one tries to represent in the present all the information needed to predict the future. With the Hamiltonian method, which begins by assuming such a representation, it is difficult to suggest modifications of a general kind, for one cannot formulate the problem without having a complete representation of the characteristics of the intermediate states, the particles involved in interaction, etc. It is quite possible (in the author's opinion, it is very likely) that we may discover that in nature the relation of past and future is so intimate for short durations that no simple representation of a present may exist. In such a case a theory could not find expression in Hamiltonian form.

An exactly similar analysis can be made just as easily starting with the general forms (46), (48). Also a coordinate representation of the photons could have been used instead of the familiar momentum one. One can deduce the rules (60), (61). Nothing essentially different is involved physically, however, so we shall not pursue the subject further here. Since they imply<sup>25</sup> all the rules for real photons, Eqs. (46), (47), (48) constitute a compact statement of all the laws of quantum electrodynamics. But they give divergent results. Can the result after charge and mass renormalization also be expressed to all orders in  $e^2/\hbar c$  in a simple way?

#### APPENDIX C. DIFFERENTIAL EQUATION FOR ELECTRON PROPAGATION

An attempt has been made to find a differential wave equation for the propagation of an electron interacting with itself, analogous to the Dirac equation, but containing terms representing the self-action. Neglecting all effects of closed loops, one such equation has been found, but not much has been done with it. It is reported here for whatever value it may have.

An electron acting upon itself is, from one point of view, a complex system of a particle and a field of an indefinite number of photons. To find a differential law of propagation of such a system we must ask first what quantities known at one instant will permit the calculation of these same quantities an instant later. Clearly, a knowledge of the position of the particle is not enough. We should need to specify: (1) the amplitude that the electron is at  $x$  and there are no photons in the field, (2) the amplitude the electron is at  $x$  and there is one photon of such and such a kind in the field, (3) the amplitude there are two photons, etc. That is, a series of functions of ever increasing numbers of variables. Following this view, we shall be led to the wave equation of the theory of second quantization.

We may also take a different view. Suppose we know a quantity  $\Phi_e[B, x]$ , a spinor function of  $x_\mu$  and functional of  $B_\mu(1)$ , defined as the amplitude that an electron arrives at  $x_\mu$  with no photon in the field when it moves in an arbitrary external unquantized potential  $B_\mu(1)$ . We allow the electron also to interact with itself,

but  $\Phi_{e2}$  is the amplitude at a given instant that there happens to be no photons present. As we have seen, a complete knowledge of this functional will also tell us the amplitude that the electron arrives at  $x$  and there is just one photon, of form  $A_\mu^{PH}(1)$  present. It is, from (60),  $\int (\delta\Phi_{e2}[B, x]/\delta B_\mu(1)) A_\mu^{PH}(1) d\tau_1$ .

Higher numbers of photons correspond to higher functional derivatives of  $\Phi_{e2}$ . Therefore,  $\Phi_{e2}[B, x]$  contains all the information requisite for describing the state of the electron-photon system, and we may expect to find a differential equation for it. Actually it satisfies  $(\nabla = \gamma_\mu \partial/\partial x_\mu, B = \gamma_\mu B_\mu)$ ,

$$(i\nabla - m)\Phi_{e2}[B, x] = B(x)\Phi_{e2}[B, x] + ie^2\gamma_\mu \int \delta_+(s_{x1}^2)(\delta\Phi_{e2}[B, x]/\delta B_\mu(1)) d\tau_1 \quad (1C)$$

as may be seen from a physical argument.<sup>25</sup> The operator  $(i\nabla - m)$  operating on the  $x$  coordinate of  $\Phi_{e2}$  should equal, from Dirac's equation, the changes in  $\Phi_{e2}$  as we go from one position  $x$  to a neighboring position due to the action of vector potentials. The term  $B(x)\Phi_{e2}$  is the effect of the external potential. But  $\Phi_{e2}$  may

<sup>25</sup> Its general validity can also be demonstrated mathematically from (45). The amplitude for arriving at  $x$  with no photons in the field with virtual photon coupling  $e^2$  is a transition amplitude. It must, therefore, satisfy (45) with  $T_{e2}[B] = \Phi_{e2}[B, x]$  for any  $x$ . Hence show that the quantity

$$C_{e2}[B, x] = (i\nabla - m - B(x))\Phi_{e2}[B, x] - ie^2\gamma_\mu \int \delta_+(s_{x1}^2)(\delta\Phi_{e2}[B, x]/\delta B_\mu(1)) d\tau_1$$

also satisfies Eq. (45) by substituting  $C_{e2}[B, x]$  for  $T_{e2}[B]$  in (45) and using the fact that  $\Phi_{e2}[B, x]$  satisfies (45). Hence if  $C_0[B, x] = 0$  then  $C_{e2}[B, x] = 0$  for all  $e^2$ . But  $C_{e2}[B, x] = 0$  means that  $\Phi_{e2}[B, x]$  satisfies (1C). Therefore, that solution  $\Phi_{e2}[B, x]$  of (45) which also satisfies  $(i\nabla - m - B(x))\Phi_0[B, x] = 0$  (the propagation of a free electron without virtual photons) is a solution of (1C) as we wished to show. Equation (1C) may be more convenient than (45) for some purposes for it does not involve differentiation with respect to the coupling constant, and is more analogous to a wave equation.

also change for at the first position  $x$  we may have had a photon present (amplitude that it was emitted at another point 1 is  $\delta\Phi_{e2}/\delta B_\mu(1)$ ) which was absorbed at  $x$  (amplitude photon released at 1 gets to  $x$  is  $\delta_+(s_{x1}^2)$  where  $s_{x1}^2$  is the squared invariant distance from 1 to  $x$ ) acting as a vector potential there (factor  $\gamma_\mu$ ). Effects of vacuum polarization are left out.

Expansion of the solution of (1C) in a power series in  $B$  and  $e^2$  starting from a free particle solution for a single electron, produces a series of terms which agree with the rules of II for action of potentials and virtual photons to various orders. It is another matter to use such an equation for the practical solution of a problem to all orders in  $e^2$ . It might be possible to represent the self-energy problem as the variational problem for  $m$ , stemming from (1C). The  $\delta_+$  will first have to be modified to obtain a convergent result.

We are not in need of the general solution of (1C). (In fact, we have it in (46), (48) in terms of the solution  $T_0[B] = \Phi_0[B, x]$  of the ordinary Dirac equation  $(i\nabla - m)\Phi_0[B, x] = B\Phi_0[B, x]$ . The general solution is too complicated, for complete knowledge of the motion of a self-acting electron in an arbitrary potential is essentially all of electrodynamics (because of the kind of relation of real and virtual processes discussed for photons in Appendix B, extended also to real and virtual pairs). Furthermore, it is easy to see that other quantities also satisfy (1C). Consider a system of many electrons, and single out some one for consideration, supposing all the others go from some definite initial state  $i$  to some definite final state  $f$ . Let  $\Phi_{e2}[B, x]$  be the amplitude that the special electron arrives at  $x$ , there are no photons present, and the other electrons go from  $i$  to  $f$  when there is an external potential  $B_\mu$  present (which  $B_\mu$  also acts on the other electrons). Then  $\Phi_{e2}$  also satisfies (1C). Likewise the amplitude with closed loops (all other electrons go vacuum to vacuum) also satisfies (1C) including all vacuum polarization effects. The various problems correspond to different assumptions as to the dependence of  $\Phi_{e2}[B, x]$  on  $B_\mu$  in the limit of zero  $e^2$ . The Eq. (1C) without further boundary conditions is probably too general to be useful.

## Equation of State Calculations by Fast Computing Machines

NICHOLAS METROPOLIS, ARIANNA W. ROSENBLUTH, MARSHALL N. ROSENBLUTH, AND AUGUSTA H. TELLER,  
*Los Alamos Scientific Laboratory, Los Alamos, New Mexico*

AND

EDWARD TELLER,\* *Department of Physics, University of Chicago, Chicago, Illinois*

(Received March 6, 1953)

A general method, suitable for fast computing machines, for investigating such properties as equations of state for substances consisting of interacting individual molecules is described. The method consists of a modified Monte Carlo integration over configuration space. Results for the two-dimensional rigid-sphere system have been obtained on the Los Alamos MANIAC and are presented here. These results are compared to the free volume equation of state and to a four-term virial coefficient expansion.

### I. INTRODUCTION

THE purpose of this paper is to describe a general method, suitable for fast electronic computing machines, of calculating the properties of any substance which may be considered as composed of interacting individual molecules. Classical statistics is assumed, only two-body forces are considered, and the potential field of a molecule is assumed spherically symmetric. These are the usual assumptions made in theories of liquids. Subject to the above assumptions, the method is not restricted to any range of temperature or density. This paper will also present results of a preliminary two-dimensional calculation for the rigid-sphere system. Work on the two-dimensional case with a Lennard-Jones potential is in progress and will be reported in a later paper. Also, the problem in three dimensions is being investigated.

### II. THE GENERAL METHOD FOR AN ARBITRARY POTENTIAL BETWEEN THE PARTICLES

In order to reduce the problem to a feasible size for numerical work, we can, of course, consider only a finite number of particles. This number  $N$  may be as high as several hundred. Our system consists of a square† containing  $N$  particles. In order to minimize the surface effects we suppose the complete substance to be periodic, consisting of many such squares, each square containing  $N$  particles in the same configuration. Thus we define  $d_{AB}$ , the minimum distance between particles  $A$  and  $B$ , as the shortest distance between  $A$  and any of the particles  $B$ , of which there is one in each of the squares which comprise the complete substance. If we have a potential which falls off rapidly with distance, there will be at most one of the distances  $AB$  which can make a substantial contribution; hence we need consider only the minimum distance  $d_{AB}$ .

\* Now at the Radiation Laboratory of the University of California, Livermore, California.

† We will use the two-dimensional nomenclature here since it is easier to visualize. The extension to three dimensions is obvious.

Our method in this respect is similar to the cell method except that our cells contain several hundred particles instead of one. One would think that such a sample would be quite adequate for describing any one-phase system. We do find, however, that in two-phase systems the surface between the phases makes quite a perturbation. Also, statistical fluctuations may be sizable.

If we know the positions of the  $N$  particles in the square, we can easily calculate, for example, the potential energy of the system,

$$E = \frac{1}{2} \sum_{\substack{i=1 \\ i \neq j}}^N \sum_{j=1}^N V(d_{ij}). \quad (1)$$

(Here  $V$  is the potential between molecules, and  $d_{ij}$  is the minimum distance between particles  $i$  and  $j$  as defined above.)

In order to calculate the properties of our system we use the canonical ensemble. So, to calculate the equilibrium value of any quantity of interest  $F$ ,

$$\bar{F} = \left[ \int F \exp(-E/kT) d^{2N} p d^{2N} q \right] / \left[ \int \exp(-E/kT) d^{2N} p d^{2N} q \right], \quad (2)$$

where  $(d^{2N} p d^{2N} q)$  is a volume element in the  $4N$ -dimensional phase space. Moreover, since forces between particles are velocity-independent, the momentum integrals may be separated off, and we need perform only the integration over the  $2N$ -dimensional configuration space. It is evidently impractical to carry out a several hundred-dimensional integral by the usual numerical methods, so we resort to the Monte Carlo method.<sup>‡</sup> The Monte Carlo method for many-dimensional integrals consists simply of integrating over a random sampling of points instead of over a regular array of points.

Thus the most naive method of carrying out the integration would be to put each of the  $N$  particles at a random position in the square (this defines a random point in the  $2N$ -dimensional configuration space), then calculate the energy of the system according to Eq. (1), and give this configuration a weight  $\exp(-E/kT)$ . This method, however, is not practical for close-packed configurations, since with high probability we choose a configuration where  $\exp(-E/kT)$  is very small; hence a configuration of very low weight. So the method we employ is actually a modified Monte Carlo scheme, where, instead of choosing configurations randomly, then weighting them with  $\exp(-E/kT)$ , we choose

<sup>‡</sup> This method has been proposed independently by J. E. Mayer and by S. Ulam. Mayer suggested the method as a tool to deal with the problem of the liquid state, while Ulam proposed it as a procedure of general usefulness. B. Alder, J. Kirkwood, S. Frankel, and V. Lewinson discussed an application very similar to ours.

configurations with a probability  $\exp(-E/kT)$  and weight them evenly.

This we do as follows: We place the  $N$  particles in any configuration, for example, in a regular lattice. Then we move each of the particles in succession according to the following prescription:

$$\begin{aligned} X &\rightarrow X + \alpha \xi_1 \\ Y &\rightarrow Y + \alpha \xi_2, \end{aligned} \quad (3)$$

where  $\alpha$  is the maximum allowed displacement, which for the sake of this argument is arbitrary, and  $\xi_1$  and  $\xi_2$  are random numbers<sup>§</sup> between  $(-1)$  and  $1$ . Then, after we move a particle, it is equally likely to be anywhere within a square of side  $2\alpha$  centered about its original position. (In accord with the periodicity assumption, if the indicated move would put the particle outside the square, this only means that it re-enters the square from the opposite side.)

We then calculate the change in energy of the system  $\Delta E$ , which is caused by the move. If  $\Delta E < 0$ , i.e., if the move would bring the system to a state of lower energy, we allow the move and put the particle in its new position. If  $\Delta E > 0$ , we allow the move with probability  $\exp(-\Delta E/kT)$ ; i.e., we take a random number  $\xi_3$  between  $0$  and  $1$ , and if  $\xi_3 < \exp(-\Delta E/kT)$ , we move the particle to its new position. If  $\xi_3 > \exp(-\Delta E/kT)$ , we return it to its old position. Then, whether the move has been allowed or not, i.e., whether we are in a different configuration or in the original configuration, we consider that we are in a new configuration for the purpose of taking our averages. So

$$\bar{F} = (1/M) \sum_{j=1}^M F_j, \quad (4)$$

where  $F_j$  is the value of the property  $F$  of the system after the  $j$ th move is carried out according to the complete prescription above. Having attempted to move a particle we proceed similarly with the next one.

We now prove that the method outlined above does choose configurations with a probability  $\exp(-E/kT)$ . Since a particle is allowed to move to any point within a square of side  $2\alpha$  with a finite probability, it is clear that a large enough number of moves will enable it to reach any point in the complete square.<sup>||</sup> Since this is true of all particles, we may reach any point in configuration space. Hence, the method is ergodic.

Next consider a very large ensemble of systems. Suppose for simplicity that there are only a finite number of states<sup>¶</sup> of the system, and that  $v_r$  is the number of

<sup>§</sup> It might be mentioned that the random numbers that we used were generated by the middle square process. That is, if  $\xi^n$  is an  $m$  digit random number, then a new random number  $\xi^{n+1}$  is given as the middle  $m$  digits of the complete  $2m$  digit square of  $\xi^n$ .

<sup>||</sup> In practice it is, of course, not necessary to make enough moves to allow a particle to diffuse evenly throughout the system since configuration space is symmetric with respect to interchange of particles.

<sup>¶</sup> A state here means a given point in configuration space.

systems of the ensemble in state  $r$ . What we must prove is that after many moves the ensemble tends to a distribution

$$\nu_r \propto \exp(-E_r/kT).$$

Now let us make a move in all the systems of our ensemble. Let the *a priori* probability that the move will carry a system in state  $r$  to state  $s$  be  $P_{rs}$ . [By the *a priori* probability we mean the probability before discriminating on  $\exp(-\Delta E/kT)$ .] First, it is clear that  $P_{rs} = P_{sr}$ , since according to the way our game is played a particle is equally likely to be moved anywhere within a square of side  $2\alpha$  centered about its original position. Thus, if states  $r$  and  $s$  differ from each other only by the position of the particle moved and if these positions are within each other's squares, the transition probabilities are equal; otherwise they are zero. Assume  $E_r > E_s$ . Then the number of systems moving from state  $r$  to state  $s$  will be simply  $\nu_r P_{rs}$ , since all moves to a state of lower energy are allowed. The number moving from  $s$  to  $r$  will be  $\nu_s P_{sr} \exp(-(E_r - E_s)/kT)$ , since here we must weigh by the exponential factor. Thus the net number of systems moving from  $s$  to  $r$  is

$$\nu_{rs} (\nu_s \exp(-(E_r - E_s)/kT) - \nu_r). \quad (5)$$

So we see that between any two states  $r$  and  $s$ , if

$$(\nu_r / \nu_s) > [\exp(-E_r/kT) / \exp(-E_s/kT)], \quad (6)$$

on the average more systems move from state  $r$  to state  $s$ . We have seen already that the method is ergodic; i.e., that any state can be reached from any other, albeit in several moves. These two facts mean that our ensemble must approach the canonical distribution. It is, incidentally, clear from the above derivation that after a forbidden move we must count again the initial configuration. Not to do this would correspond in the above case to removing from the ensemble those systems which tried to move from  $s$  to  $r$  and were forbidden. This would unjustifiably reduce the number in state  $s$  relative to  $r$ .

The above argument does not, of course, specify how rapidly the canonical distribution is approached. It may be mentioned in this connection that the maximum displacement  $\alpha$  must be chosen with some care; if too large, most moves will be forbidden, and if too small, the configuration will not change enough. In either case it will then take longer to come to equilibrium.

For the rigid-sphere case, the game of chance on  $\exp(-\Delta E/kT)$  is, of course, not necessary since  $\Delta E$  is either zero or infinity. The particles are moved, one at a time, according to Eq. (3). If a sphere, after such a move, happens to overlap another sphere, we return it to its original position.

### III. SPECIALIZATION TO RIGID SPHERES IN TWO DIMENSIONS

#### A. The Equation of State

The virial theorem of Clausius can be used to give an equation of state in terms of  $\bar{n}$ , the average den-

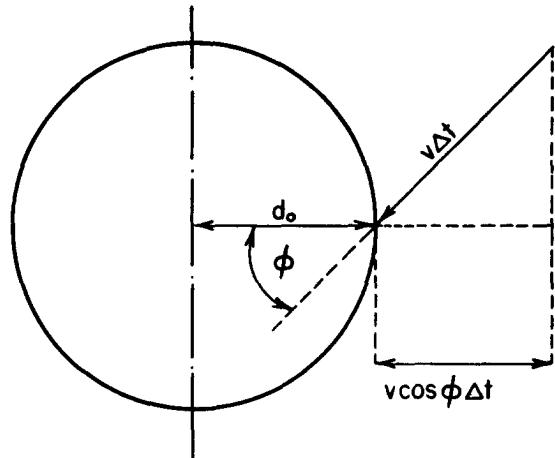


FIG. 1. Collisions of rigid spheres.

sity of other particles at the surface of a particle. Let  $\mathbf{X}_i^{(tot)}$  and  $\mathbf{X}_i^{(int)}$  represent the total and the internal force, respectively, acting on particle  $i$ , at a position  $\mathbf{r}_i$ . Then the virial theorem can be written

$$\left\langle \sum_i \mathbf{X}_i^{(tot)} \cdot \mathbf{r}_i \right\rangle_{Av} = 2PA + \left\langle \sum_i \mathbf{X}_i^{(int)} \cdot \mathbf{r}_i \right\rangle_{Av} = 2E_{kin}. \quad (7)$$

Here  $P$  is the pressure,  $A$  the area, and  $E_{kin}$  the total kinetic energy,

$$E_{kin} = Nm\bar{v}^2/2$$

of the system of  $N$  particles.

Consider the collisions of the spheres for convenience as represented by those of a particle of radius  $d_0$ , twice the radius of the actual spheres, surrounded by  $\bar{n}$  point particles per unit area. Those surrounding particles in an area of  $2\pi d_0 \cos \phi \Delta t$ , traveling with velocity  $v$  at an angle  $\phi$  with the radius vector, collide with the central particle provided  $|\phi| < \pi/2$ . (See Fig. 1.) Assuming elastic recoil, they each exert an average force during the time  $\Delta t$  on the central particle of

$$2mv \cos \phi / \Delta t.$$

One can see that all  $\phi$ 's are equally probable, since for any velocity-independent potential between particles the velocity distribution will just be Maxwellian, hence isotropic. The total force acting on the central particle, averaged over  $\phi$ , over time, and over velocity, is

$$\bar{F}_i = m\bar{v}^2 \pi d_0 \bar{n}. \quad (8)$$

The sum

$$\left\langle \sum_i \mathbf{X}_i^{(int)} \cdot \mathbf{r}_i \right\rangle_{Av}$$

is

$$-\frac{1}{2} \sum_i \sum_{j \neq i} r_{ij} F_{ij},$$

with  $F_{ij}$  the magnitude of the force between two particles and  $r_{ij}$  the distance between them. We see that

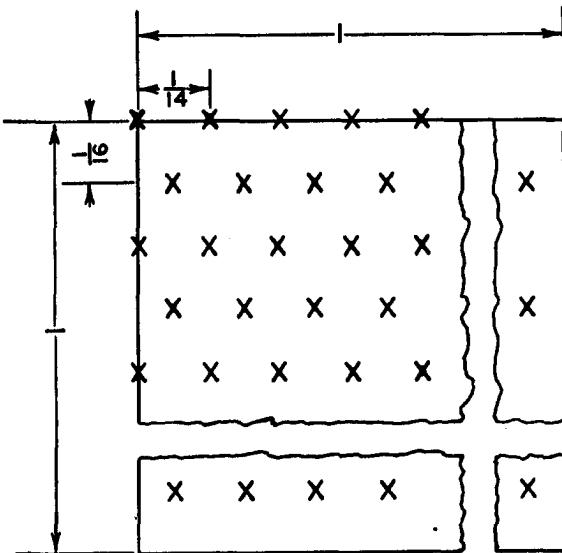


FIG. 2. Initial trigonal lattice.

$r_{ij} = d_0$  and  $\sum_j F_{ij}$  is given by Eq. (8), so we have

$$\langle \sum_i \mathbf{X}_i^{(\text{int})} \cdot \mathbf{r}_i \rangle_A = -(Nm\bar{v}^2/2)\pi d_0^2 \bar{n}. \quad (9)$$

Substitution of (9) into (7) and replacement of  $(N/2)m\bar{v}^2$  by  $E_{\text{kin}}$  gives finally

$$PA = E_{\text{kin}}(1 + \pi d_0^2 \bar{n}/2) \equiv NkT(1 + \pi d_0^2 \bar{n}/2). \quad (10)$$

This equation shows that a determination of the one quantity  $\bar{n}$ , according to Eq. (4) as a function of  $A$ , the area, is sufficient to determine the equation of state for the rigid spheres.

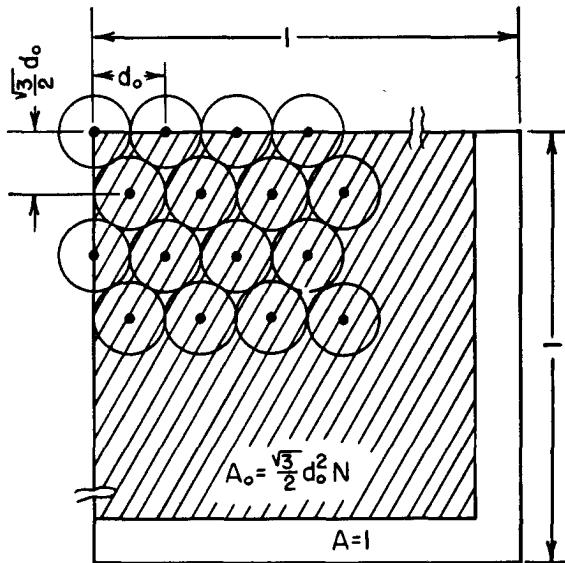
### B. The Actual Calculation of $\bar{n}$

We set up the calculation on a system composed of  $N = 224$  particles ( $i = 0, 1 \dots 223$ ) placed inside a square of unit side and unit area. The particles were arranged initially in a trigonal lattice of fourteen particles per row by sixteen particles per column, alternate rows being displaced relative to each other as shown in Fig. 2. This arrangement gives each particle six nearest neighbors at approximately equal distances of  $d = 1/14$  from it.

Instead of performing the calculation for various areas  $A$  and for a fixed distance  $d_0$ , we shall solve the equivalent problem of leaving  $A = 1$  fixed and changing  $d_0$ . We denote by  $A_0$  the area the particles occupy in close-packed arrangement (see Fig. 3). For numerical convenience we defined an auxiliary parameter  $\nu$ , which we varied from zero to seven, and in terms of which the ratio  $(A/A_0)$  and the forbidden distance  $d_0$  are defined as follows:

$$d_0 = d(1 - 2^{-\nu}), \quad d = (1/14), \quad (11a)$$

$$(A/A_0) = 1/(3^{\frac{1}{2}} d_0^2 N/2) = 1/0.98974329(1 - 2^{-\nu})^2. \quad (11b)$$

FIG. 3. The close-packed arrangement for determining  $A_0$ .

The unit cell is a parallelogram with interior angle  $60^\circ$ , side  $d_0$ , and altitude  $3^{\frac{1}{2}}d_0/2$  in the close-packed system.

Every configuration reached by proceeding according to the method of the preceding section was analyzed in terms of a radial distribution function  $N(r^2)$ . We chose a  $K > 1$  for each  $\nu$  and divided the area between  $\pi d_0^2$  and  $K^2 \pi d_0^2$  into sixty-four zones of equal area  $\Delta A^2$ ,

$$\Delta A^2 = (K^2 - 1)\pi d_0^2 / 64.$$

We then had the machine calculate for each configuration the number of pairs of particles  $N_m$  ( $m = 1, 2, \dots, 64$ ) separated by distances  $r$  which satisfy

$$(m-1)\Delta A^2 + \pi d_0^2 < \pi r^2 \leq m\Delta A^2 + \pi d_0^2. \quad (12)$$

The  $N_m$  were averaged over successive configurations according to Eq. (4), and after every sixteen cycles (a cycle consists of moving every particle once) were extrapolated back to  $r^2 = d_0^2$  to obtain  $N_{\frac{1}{2}}$ . This  $N_{\frac{1}{2}}$  differs from  $\bar{n}$  in Eq. (10) by a constant factor depending on  $N$  and  $K$ .

The quantity  $K$  was chosen for each  $\nu$  to give reasonable statistics for the  $N_m$ . It would, of course, have been possible by choosing fairly large  $K$ 's, with perhaps a larger number of zones, to obtain  $N(r^2)$  at large distances. The oscillatory behavior of  $N(r^2)$  at large distances is of some interest. However, the time per cycle goes up fairly rapidly with  $K$  and with the number of zones in the distance analysis. For this reason only the behavior of  $N(r^2)$  in the neighborhood of  $d_0^2$  was investigated.

The maximum displacement  $\alpha$  of Eq. (3) was set to  $(d - d_0)$ . About half the moves in a cycle were forbidden by this choice, and the initial approach to equilibrium from the regular lattice was fairly rapid.

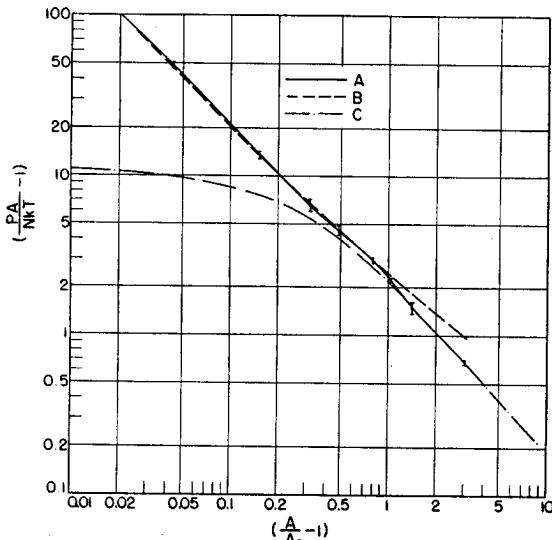


FIG. 4. A plot of  $(PA/NkT)-1$  versus  $(A/A_0)-1$ . Curve A (solid line) gives the results of this paper. Curves B and C (dashed and dot-dashed lines) give the results of the free volume theory and of the first four virial coefficients, respectively.

#### IV. NUMERICAL RESULTS FOR RIGID SPHERES IN TWO DIMENSIONS

We first ran for something less than sixteen cycles in order to get rid of the effects of the initial regular configuration on the averages. Then about forty-eight to sixty-four cycles were run at

$$\nu = 2, 4, 5, 5.5, 6, 6.25, 6.5, \text{ and } 7.$$

Also, a smaller amount of data was obtained at  $\nu=0, 1$ , and 3. The time per cycle on the Los Alamos MANIAC is approximately three minutes, and a given point on the pressure curve was obtained in four to five hours of running. Figure 4 shows  $(PA/NkT)-1$  versus  $(A/A_0)-1$  on a log-log scale from our results (curve A), compared to the free volume equation of Wood<sup>1</sup> (curve B) and to the curve given by the first four virial coefficients (curve C). The last two virial coefficients were obtained by straightforward Monte Carlo integration on the MANIAC (see Sec. V). It is seen that the agreement between curves A and B at small areas and between curves A and C at large areas is good. Deviation from the free volume theory begins with a fairly sudden break at  $\nu=6$  ( $A/A_0 \approx 1.8$ ).

A sample plot of the radial distribution function for  $\nu=5$  is given in Fig. 5. The various types of points represent values after sixteen, thirty-two, and forty-eight cycles. For  $\nu=5$ , least-square fits with a straight line to the first sixteen  $N_m$  values were made, giving extrapolated values of  $N_{\frac{1}{2}}^{(1)}=6367$ ,  $N_{\frac{1}{2}}^{(2)}=6160$ , and  $N_{\frac{1}{2}}^{(3)}=6377$ . The average of these three was used in constructing  $PA/NkT$ . In general, least-square fits of the first sixteen to twenty  $N_m$ 's by means of a parabola, or, where it seemed suitable, a straight line, were made.

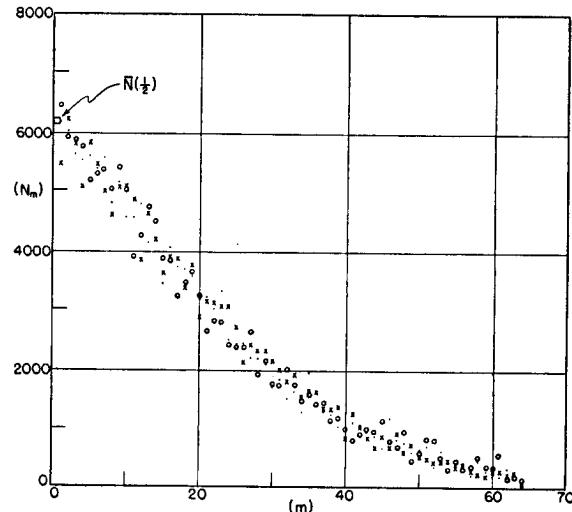


FIG. 5. The radial distribution function  $N_m$  for  $\nu=5$ ,  $(A/A_0)=1.31966$ ,  $K=1.5$ . The average of the extrapolated values of  $N_{\frac{1}{2}}$  in  $N_{\frac{1}{2}}=6301$ . The resultant value of  $(PA/NkT)-1$  is  $64N_{\frac{1}{2}}/N^2(K^2-1)$  or 6.43. Values after 16 cycles, ●; after 32, X; and after 48, ○.

The errors indicated in Fig. 4 are the root-mean-square deviations for the three or four  $N_{\frac{1}{2}}$  values. Our average error seemed to be about 3 percent.

Table I gives the results of our calculations in numerical form. The columns are  $\nu$ ,  $A/A_0$ ,  $(PA/NkT)-1$ , and, for comparison purposes,  $(PA/NkT-1)$  for the free volume theory and for the first four coefficients in the virial coefficient expansion, in that order, and finally  $PA_0/NkT$  from our results.

#### V. THE VIRIAL COEFFICIENT EXPANSION

One can show<sup>2</sup> that

$$(PA/NkT)-1 = C_1(A_0/A) + C_2(A_0/A)^2 + C_3(A_0/A)^3 + C_4(A_0/A)^4 + O(A_0/A)^5,$$

$$C_1 = \pi/3^{\frac{1}{2}}, \quad C_2 = 4\pi^2 A_{3,3}/9,$$

$$C_3 = \pi^3 (6A_{4,5} - 3A_{4,4} - A_{4,6})/3^{\frac{1}{2}},$$

$$C_4 = (8\pi^3/135) \cdot [12A_{5,5} - 60A_{5,6}' - 10A_{5,6}'' + 30A_{5,7}' + 60A_{5,7}'' + 10A_{5,7}''' - 30A_{5,8}' - 15A_{5,8}'' + 10A_{5,9} - A_{5,10}]. \quad (13)$$

TABLE I. Results of this calculation for  $(PA/NkT)-1=X_1$  compared to the free volume theory ( $X_2$ ) and the four-term virial expansion ( $X_3$ ). Also  $(PA_0/NkT)$  from our calculations.

$\nu$	$(A/A_0)$	$X_1$	$X_2$	$X_3$	$(PA_0/NkT)$
2	1.04269	49.17	47.35	9.77	48.11
4	1.14957	13.95	13.85	7.55	13.01
5	1.31966	6.43	6.72	5.35	5.63
5.5	1.4909	4.41	4.53	4.02	3.63
6	1.7962	2.929	2.939	2.680	2.187
6.25	2.04616	2.186	2.323	2.065	1.557
6.5	2.41751	1.486	1.802	1.514	1.028
7	4.04145	0.6766	0.990	0.667	0.4149

<sup>2</sup> J. E. Mayer and M. G. Mayer, *Statistical Mechanics* (John Wiley and Sons, Inc., New York, 1940), pp. 277-291.

<sup>1</sup> William W. Wood, *J. Chem. Phys.* **20**, 1334 (1952).

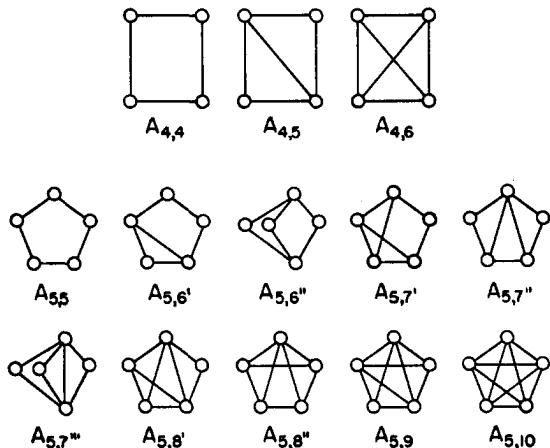


FIG. 6. Schematic diagrams for the various area integrals.

The coefficients  $A_{i,k}$  are cluster integrals over configuration space of  $i$  particles, with  $k$  bonds between them. In our problem a bond is established if the two particles overlap. The cluster integral is the volume of configuration space for which the appropriate bonds are established. If  $k$  bonds can be distributed over the  $i$  particles in two or more different ways without destroying the irreducibility of the integrals, the separate cases are distinguished by primes. For example,  $A_{3,3}$  is given schematically by the diagram



and mathematically as follows: if we define  $f(r_{ij})$  by

$$\begin{aligned} f(r_{ij}) &= 1 \quad \text{if } r_{ij} < d, \\ f(r_{ij}) &= 0 \quad \text{if } r_{ij} > d, \end{aligned}$$

then

$$A_{3,3} = \frac{1}{\pi^2 d^4} \int \cdots \int dx_1 dx_2 dx_3 dy_1 dy_2 dy_3 (f_{12} f_{23} f_{31}).$$

The schematics for the remaining integrals are indicated in Fig. 6.

The coefficients  $A_{3,3}$ ,  $A_{4,4}$ , and  $A_{4,5}$  were calculated algebraically, the remainder numerically by Monte Carlo integration. That is, for  $A_{5,5}$  for example, particle 1 was placed at the origin, and particles 2, 3, 4, and 5

were put down at random, subject to  $f_{12}=f_{23}=f_{34}=f_{15}=1$ . The number of trials for which  $f_{45}=1$ , divided by the total number of trials, is just  $A_{5,5}$ .

The data on  $A_{4,6}$  is quite reliable. We obtained

$$A_{4,6}/A_{4,4} = 0.752 (\pm 0.002).$$

However, because of the relatively large positive and negative terms in  $C_4$  of Eq. (13), the coefficient  $C_4$ , being a small difference, is less accurate. We obtained

$$C_4 = 8\pi^3(0.585)/135 \quad (\pm \sim 5 \text{ percent}).$$

Our final formula is

$$\begin{aligned} (PA/NkT) - 1 &= 1.813799(A_0/A) \\ &\quad + 2.57269(A_0/A)^2 + 3.179(A_0/A)^3 \\ &\quad + 3.38(A_0/A)^4 + 0(A_0/A)^5. \end{aligned} \quad (14)$$

This formula is plotted in curve C of Fig. 4 and tabulated for some values of  $(A/A_0)$  in column 5 of Table I. It is seen in Fig. 4 that the curves agree very well with our calculated equation of state for  $(A/A_0) > 2.5$ . In this region both the possible error in our last virial coefficients and the contribution of succeeding terms in the expansion are quite small (less than our probable statistical error) so that the virial expansion should be accurate.

## VI. CONCLUSION

The method of Monte Carlo integrations over configuration space seems to be a feasible approach to statistical mechanical problems which are as yet not analytically soluble. At least for a single-phase system a sample of several hundred particles seems sufficient. In the case of two-dimensional rigid spheres, runs made with 56 particles and with 224 particles agreed within statistical error. For a computing time of a few hours with presently available electronic computers, it seems possible to obtain the pressure for a given volume and temperature to an accuracy of a few percent.

In the case of two-dimensional rigid spheres our results are in agreement with the free volume approximation for  $A/A_0 < 1.8$  and with a five-term virial expansion for  $A/A_0 > 2.5$ . There is no indication of a phase transition.

Work is now in progress for a system of particles with Lennard-Jones type interactions and for three-dimensional rigid spheres.

## Conservation of Isotopic Spin and Isotopic Gauge Invariance\*

C. N. YANG † AND R. L. MILLS

Brookhaven National Laboratory, Upton, New York

(Received June 28, 1954)

It is pointed out that the usual principle of invariance under isotopic spin rotation is not consistent with the concept of localized fields. The possibility is explored of having invariance under local isotopic spin rotations. This leads to formulating a principle of isotopic gauge invariance and the existence of a **b** field which has the same relation to the isotopic spin that the electromagnetic field has to the electric charge. The **b** field satisfies nonlinear differential equations. The quanta of the **b** field are particles with spin unity, isotopic spin unity, and electric charge  $\pm e$  or zero.

### INTRODUCTION

THE conservation of isotopic spin is a much discussed concept in recent years. Historically an isotopic spin parameter was first introduced by Heisenberg<sup>1</sup> in 1932 to describe the two charge states (namely neutron and proton) of a nucleon. The idea that the neutron and proton correspond to two states of the same particle was suggested at that time by the fact that their masses are nearly equal, and that the light

stable even nuclei contain equal numbers of them. Then in 1937 Breit, Condon, and Present pointed out the approximate equality of  $p-p$  and  $n-p$  interactions in the  $^1S$  state.<sup>2</sup> It seemed natural to assume that this equality holds also in the other states available to both the  $n-p$  and  $p-p$  systems. Under such an assumption one arrives at the concept of a total isotopic spin<sup>3</sup> which is conserved in nucleon-nucleon interactions. Experi-

\* Work performed under the auspices of the U. S. Atomic Energy Commission.

† On leave of absence from the Institute for Advanced Study, Princeton, New Jersey.

<sup>1</sup> W. Heisenberg, Z. Physik **77**, 1 (1932).

<sup>2</sup> Breit, Condon, and Present, Phys. Rev. **50**, 825 (1936). J. Schwinger pointed out that the small difference may be attributed to magnetic interactions [Phys. Rev. **78**, 135 (1950)].

<sup>3</sup> The total isotopic spin **T** was first introduced by E. Wigner, Phys. Rev. **51**, 106 (1937); B. Casen and E. U. Condon, Phys. Rev. **50**, 846 (1936).

ments in recent years<sup>4</sup> on the energy levels of light nuclei strongly suggest that this assumption is indeed correct. An implication of this is that all strong interactions such as the pion-nucleon interaction, must also satisfy the same conservation law. This and the knowledge that there are three charge states of the pion, and that pions can be coupled to the nucleon field *singly*, lead to the conclusion that pions have isotopic spin unity. A direct verification of this conclusion was found in the experiment of Hildebrand<sup>5</sup> which compares the differential cross section of the process  $n+p \rightarrow \pi^0 + d$  with that of the previously measured process  $p+p \rightarrow \pi^+ + d$ .

The conservation of isotopic spin is identical with the requirement of invariance of all interactions under isotopic spin rotation. This means that when electromagnetic interactions can be neglected, as we shall hereafter assume to be the case, the orientation of the isotopic spin is of no physical significance. The differentiation between a neutron and a proton is then a purely arbitrary process. As usually conceived, however, this arbitrariness is subject to the following limitation: once one chooses what to call a proton, what a neutron, at one space-time point, one is then not free to make any choices at other space-time points.

It seems that this is not consistent with the localized field concept that underlies the usual physical theories. In the present paper we wish to explore the possibility of requiring all interactions to be invariant under *independent* rotations of the isotopic spin at all space-time points, so that the relative orientation of the isotopic spin at two space-time points becomes a physically meaningless quantity (the electromagnetic field being neglected).

We wish to point out that an entirely similar situation arises with respect to the ordinary gauge invariance of a charged field which is described by a complex wave function  $\psi$ . A change of gauge<sup>6</sup> means a change of phase factor  $\psi \rightarrow \psi'$ ,  $\psi' = (\exp i\alpha)\psi$ , a change that is devoid of any physical consequences. Since  $\psi$  may depend on  $x, y, z$ , and  $t$ , the relative phase factor of  $\psi$  at two different space-time points is therefore completely arbitrary. In other words, the arbitrariness in choosing the phase factor is local in character.

We define *isotopic gauge* as an arbitrary way of choosing the orientation of the isotopic spin axes at all space-time points, in analogy with the electromagnetic gauge which represents an arbitrary way of choosing the complex phase factor of a charged field at all space-time points. We then propose that all physical processes (not involving the electromagnetic field) be invariant under an isotopic gauge transformation,  $\psi \rightarrow \psi'$ ,  $\psi' = S^{-1}\psi$ , where  $S$  represents a space-time dependent isotopic spin rotation.

To preserve invariance one notices that in electro-

dynamics it is necessary to counteract the variation of  $\alpha$  with  $x, y, z$ , and  $t$  by introducing the electromagnetic field  $A_\mu$  which changes under a gauge transformation as

$$A'_\mu = A_\mu + \frac{1}{e} \frac{\partial \alpha}{\partial x_\mu}$$

In an entirely similar manner we introduce a  $B$  field in the case of the isotopic gauge transformation to counteract the dependence of  $S$  on  $x, y, z$ , and  $t$ . It will be seen that this natural generalization allows for very little arbitrariness. The field equations satisfied by the twelve independent components of the  $B$  field, which we shall call the  $\mathbf{b}$  field, and their interaction with any field having an isotopic spin are essentially fixed, in much the same way that the free electromagnetic field and its interaction with charged fields are essentially determined by the requirement of gauge invariance.

In the following two sections we put down the mathematical formulation of the idea of isotopic gauge invariance discussed above. We then proceed to the quantization of the field equations for the  $\mathbf{b}$  field. In the last section the properties of the quanta of the  $\mathbf{b}$  field are discussed.

#### ISOTOPIC GAUGE TRANSFORMATION

Let  $\psi$  be a two-component wave function describing a field with isotopic spin  $\frac{1}{2}$ . Under an isotopic gauge transformation it transforms by

$$\psi' = S\psi, \quad (1)$$

where  $S$  is a  $2 \times 2$  unitary matrix with determinant unity. In accordance with the discussion in the previous section, we require, in analogy with the electromagnetic case, that all derivatives of  $\psi$  appear in the following combination:

$$(\partial_\mu - i\epsilon B_\mu)\psi.$$

$B_\mu$  are  $2 \times 2$  matrices such that<sup>7</sup> for  $\mu = 1, 2$ , and  $3$ ,  $B_\mu$  is Hermitian and  $B_4$  is anti-Hermitian. Invariance requires that

$$S(\partial_\mu - i\epsilon B_\mu')\psi' = (\partial_\mu - i\epsilon B_\mu)\psi. \quad (2)$$

Combining (1) and (2), we obtain the isotopic gauge transformation on  $B_\mu$ :

$$B'_\mu = S^{-1}B_\mu S + \frac{i}{\epsilon} \frac{\partial S}{\partial x_\mu}. \quad (3)$$

The last term is similar to the gradient term in the gauge transformation of electromagnetic potentials. In analogy to the procedure of obtaining gauge invariant field strengths in the electromagnetic case, we

<sup>4</sup> T. Lauritsen, Ann. Rev. Nuclear Sci. **1**, 67 (1952); D. R. Inglis, Revs. Modern Phys. **25**, 390 (1953).

<sup>5</sup> R. H. Hildebrand, Phys. Rev. **89**, 1090 (1953).

<sup>6</sup> W. Pauli, Revs. Modern Phys. **13**, 203 (1941).

<sup>7</sup> We use the conventions  $\hbar = c = 1$ , and  $x_4 = it$ . Bold-face type refers to vectors in isotopic space, not in space-time.

define now

$$F_{\mu\nu} = \frac{\partial B_\mu}{\partial x_\nu} - \frac{\partial B_\nu}{\partial x_\mu} + i\epsilon(B_\mu B_\nu - B_\nu B_\mu). \quad (4)$$

One easily shows from (3) that

$$F'_{\mu\nu} = S^{-1}F_{\mu\nu}S \quad (5)$$

under an isotopic gauge transformation.<sup>‡</sup> Other simple functions of  $B$  than (4) do not lead to such a simple transformation property.

The above lines of thought can be applied to any field  $\psi$  with arbitrary isotopic spin. One need only use other representations  $S$  of rotations in three-dimensional space. It is reasonable to assume that different fields with the same total isotopic spin, hence belonging to the same representation  $S$ , interact with the same matrix field  $B_\mu$ . (This is analogous to the fact that the electromagnetic field interacts in the same way with any charged particle, regardless of the nature of the particle. If different fields interact with different and independent  $B$  fields, there would be more conservation laws than simply the conservation of total isotopic spin.) To find a more explicit form for the  $B$  fields and to relate the  $B_\mu$ 's corresponding to different representations  $S$ , we proceed as follows.

Equation (3) is valid for any  $S$  and its corresponding  $B_\mu$ . Now the matrix  $S^{-1}\partial S/\partial x_\mu$  appearing in (3) is a linear combination of the isotopic spin "angular momentum" matrices  $T^i$  ( $i=1, 2, 3$ ) corresponding to the isotopic spin of the  $\psi$  field we are considering. So  $B_\mu$  itself must also contain a linear combination of the matrices  $T^i$ . But any part of  $B_\mu$  in addition to this,  $\bar{B}_\mu$ , say, is a scalar or tensor combination of the  $T$ 's, and must transform by the homogeneous part of (3),  $\bar{B}'_\mu = S^{-1}\bar{B}_\mu S$ . Such a field is extraneous; it was allowed by the very general form we assumed for the  $B$  field, but is irrelevant to the question of isotopic gauge. Thus the relevant part of the  $B$  field is of the form

$$B_\mu = 2\mathbf{b}_\mu \cdot \mathbf{T}. \quad (6)$$

(Bold-face letters denote three-component vectors in isotopic space.) To relate the  $\mathbf{b}_\mu$ 's corresponding to different representations  $S$  we now consider the product representation  $S = S^{(a)}S^{(b)}$ . The  $B$  field for the combination transforms, according to (3), by

$$\begin{aligned} B'_\mu &= [S^{(b)}]^{-1}[S^{(a)}]^{-1}BS^{(a)}S^{(b)} \\ &\quad + \frac{i}{\epsilon} [S^{(a)}]^{-1} \frac{\partial S^{(a)}}{\partial x_\mu} + \frac{i}{\epsilon} [S^{(b)}]^{-1} \frac{\partial S^{(b)}}{\partial x_\mu}. \end{aligned}$$

<sup>‡</sup> Note added in proof.—It may appear that  $B_\mu$  could be introduced as an auxiliary quantity to accomplish invariance, but need not be regarded as a field variable by itself. It is to be emphasized that such a procedure violates the principle of invariance. Every quantity that is not a pure numeral (like 2, or  $M$ , or any definite representation of the  $\gamma$  matrices) should be regarded as a dynamical variable, and should be varied in the Lagrangian to yield an equation of motion. Thus the quantities  $B_\mu$  must be regarded as independent fields.

But the sum of  $B_\mu^{(a)}$  and  $B_\mu^{(b)}$ , the  $B$  fields corresponding to  $S^{(a)}$  and  $S^{(b)}$ , transforms in exactly the same way, so that

$$B_\mu = B_\mu^{(a)} + B_\mu^{(b)}$$

(plus possible terms which transform homogeneously, and hence are irrelevant and will not be included). Decomposing  $S^{(a)}S^{(b)}$  into irreducible representations, we see that the twelve-component field  $\mathbf{b}_\mu$  in Eq. (6) is the same for all representations.

To obtain the interaction between any field  $\psi$  of arbitrary isotopic spin with the  $\mathbf{b}$  field one therefore simply replaces the gradient of  $\psi$  by

$$(\partial_\mu - 2i\epsilon\mathbf{b}_\mu \cdot \mathbf{T})\psi, \quad (7)$$

where  $T^i$  ( $i=1, 2, 3$ ), as defined above, are the isotopic spin "angular momentum" matrices for the field  $\psi$ .

We remark that the nine components of  $\mathbf{b}_\mu$ ,  $\mu=1, 2, 3$  are real and the three of  $\mathbf{b}_4$  are pure imaginary. The isotopic-gauge covariant field quantities  $F_{\mu\nu}$  are expressible in terms of  $\mathbf{b}_\mu$ :

$$F_{\mu\nu} = 2\mathbf{f}_{\mu\nu} \cdot \mathbf{T}, \quad (8)$$

where

$$\mathbf{f}_{\mu\nu} = \frac{\partial \mathbf{b}_\mu}{\partial x_\nu} - \frac{\partial \mathbf{b}_\nu}{\partial x_\mu} - 2\epsilon\mathbf{b}_\mu \times \mathbf{b}_\nu. \quad (9)$$

$\mathbf{f}_{\mu\nu}$  transforms like a vector under an isotopic gauge transformation. Obviously the same  $\mathbf{f}_{\mu\nu}$  interact with all fields  $\psi$  irrespective of the representation  $S$  that  $\psi$  belongs to.

The corresponding transformation of  $\mathbf{b}_\mu$  is cumbersome. One need, however, study only the infinitesimal isotopic gauge transformations,

$$S = 1 - 2iT \cdot \delta\omega.$$

Then

$$\mathbf{b}'_\mu = \mathbf{b}_\mu + 2\mathbf{b}_\mu \times \delta\omega + \frac{1}{\epsilon} \frac{\partial}{\partial x_\mu} \delta\omega. \quad (10)$$

#### FIELD EQUATIONS

To write down the field equations for the  $\mathbf{b}$  field we clearly only want to use isotopic gauge invariant quantities. In analogy with the electromagnetic case we therefore write down the following Lagrangian density:<sup>§</sup>

$$-\frac{1}{4}\mathbf{f}_{\mu\nu} \cdot \mathbf{f}_{\mu\nu}.$$

Since the inclusion of a field with isotopic spin  $\frac{1}{2}$  is illustrative, and does not complicate matters very much, we shall use the following total Lagrangian density:

$$\mathcal{L} = -\frac{1}{4}\mathbf{f}_{\mu\nu} \cdot \mathbf{f}_{\mu\nu} - \bar{\psi}\gamma_\mu(\partial_\mu - i\epsilon\tau \cdot \mathbf{b}_\mu)\psi - m\bar{\psi}\psi. \quad (11)$$

One obtains from this the following equations of motion:

$$\begin{aligned} \partial\mathbf{f}_{\mu\nu}/\partial x_\nu + 2\epsilon(\mathbf{b}_\nu \times \mathbf{f}_{\mu\nu}) + \mathbf{J}_\mu &= 0, \\ \gamma_\mu(\partial_\mu - i\epsilon\tau \cdot \mathbf{b}_\mu)\psi + m\psi &= 0, \end{aligned} \quad (12)$$

<sup>§</sup> Repeated indices are summed over, except where explicitly stated otherwise. Latin indices are summed from 1 to 3, Greek ones from 1 to 4.

where

$$\mathbf{J}_\mu = i\epsilon\bar{\psi}\gamma_\mu\tau\psi. \quad (13)$$

The divergence of  $\mathbf{J}_\mu$  does not vanish. Instead it can easily be shown from (13) that

$$\partial\mathbf{J}_\mu/\partial x_\mu = -2\epsilon\mathbf{b}_\nu \times \mathbf{J}_\nu. \quad (14)$$

If we define, however,

$$\mathfrak{J}_\mu = \mathbf{J}_\mu + 2\epsilon\mathbf{b}_\nu \times \mathbf{f}_{\mu\nu}, \quad (15)$$

then (12) leads to the equation of continuity,

$$\partial\mathfrak{J}_\mu/\partial x_\mu = 0. \quad (16)$$

$\mathfrak{J}_{1,2,3}$  and  $\mathfrak{J}_4$  are respectively the isotopic spin current density and isotopic spin density of the system. The equation of continuity guarantees that the total isotopic spin

$$\mathbf{T} = \int \mathfrak{J}_4 d^3x$$

is independent of time and independent of a Lorentz transformation. It is important to notice that  $\mathfrak{J}_\mu$ , like  $\mathbf{b}_\mu$ , does not transform exactly like vectors under isotopic space rotations. But the total isotopic spin,

$$\mathbf{T} = - \int \frac{\partial \mathbf{f}_{4i}}{\partial x_i} d^3x,$$

is the integral of the divergence of  $\mathbf{f}_{4i}$ , which transforms like a true vector under isotopic spin space rotations. Hence, under a general isotopic gauge transformation, if  $S \rightarrow S_0$  on an infinitely large sphere,  $\mathbf{T}$  would transform like an isotopic spin vector.

Equation (15) shows that the isotopic spin arises both from the spin- $\frac{1}{2}$  field ( $\mathbf{J}_\mu$ ) and from the  $\mathbf{b}_\mu$  field itself. Inasmuch as the isotopic spin is the source of the  $\mathbf{b}$  field, this fact makes the field equations for the  $\mathbf{b}$  field nonlinear, even in the absence of the spin- $\frac{1}{2}$  field. This is different from the case of the electromagnetic field, which is itself chargeless, and consequently satisfies linear equations in the absence of a charged field.

The Hamiltonian derived from (11) is easily demonstrated to be positive definite in the absence of the field of isotopic spin  $\frac{1}{2}$ . The demonstration is completely identical with the similar one in electrodynamics.

We must complete the set of equations of motion (12) and (13) by the supplementary condition,

$$\partial\mathbf{b}_\mu/\partial x_\mu = 0, \quad (17)$$

which serves to eliminate the scalar part of the field in  $\mathbf{b}_\mu$ . This clearly imposes a condition on the possible isotopic gauge transformations. That is, the infinitesimal isotopic gauge transformation  $S = 1 - i\tau \cdot \delta\omega$  must satisfy the following condition:

$$2\mathbf{b}_\mu \times \frac{\partial}{\partial x_\mu} \delta\omega + \frac{1}{\epsilon} \frac{\partial^2}{\partial x_\mu^2} \delta\omega = 0. \quad (18)$$

This is the analog of the equation  $\partial^2\alpha/\partial x_\mu^2 = 0$  that must be satisfied by the gauge transformation  $A'_\mu = A_\mu + \epsilon^{-1}(\partial\alpha/\partial x_\mu)$  of the electromagnetic field.

### QUANTIZATION

To quantize, it is not convenient to use the isotopic gauge invariant Lagrangian density (11). This is quite similar to the corresponding situation in electrodynamics and we adopt the customary procedure of using a Lagrangian density which is not obviously gauge invariant:

$$\mathcal{L} = -\frac{1}{2} \frac{\partial\mathbf{b}_\mu}{\partial x_\nu} \cdot \frac{\partial\mathbf{b}_\mu}{\partial x_\nu} + 2\epsilon(\mathbf{b}_\mu \times \mathbf{b}_\nu) \frac{\partial\mathbf{b}_\mu}{\partial x_\nu} - \epsilon^2(\mathbf{b}_\mu \times \mathbf{b}_\nu)^2 + \mathbf{J}_\mu \cdot \mathbf{b}_\mu - \bar{\psi}(\gamma_\mu \partial_\mu + m)\psi. \quad (19)$$

The equations of motion that result from this Lagrangian density can be easily shown to imply that

$$\frac{\partial^2}{\partial x_\nu^2} \mathbf{a} + 2\epsilon\mathbf{b}_\nu \times \frac{\partial}{\partial x_\nu} \mathbf{a} = 0,$$

where

$$\mathbf{a} = \partial\mathbf{b}_\mu/\partial x_\mu.$$

Thus if, consistent with (17), we put on one space-like surface  $\mathbf{a} = 0$  together with  $\partial\mathbf{a}/\partial t = 0$ , it follows that  $\mathbf{a} = 0$  at all times. Using this supplementary condition one can easily prove that the field equations resulting from the Lagrangian densities (19) and (11) are identical.

One can follow the canonical method of quantization with the Lagrangian density (19). Defining

$$\mathbf{I}_\mu = -\partial\mathbf{b}_\mu/\partial x_4 + 2\epsilon(\mathbf{b}_\mu \times \mathbf{b}_4),$$

one obtains the equal-time commutation rule

$$[b_\mu^i(x), \mathbf{I}_\nu^j(x')]_{t=t'} = -\delta_{ij}\delta_{\mu\nu}\delta^3(x-x'), \quad (20)$$

where  $b_\mu^i$ ,  $i=1, 2, 3$ , are the three components of  $\mathbf{b}_\mu$ . The relativistic invariance of these commutation rules follows from the general proof for canonical methods of quantization given by Heisenberg and Pauli.<sup>9</sup>

The Hamiltonian derived from (19) is identical with the one from (11), in virtue of the supplementary condition. Its density is

$$H = H_0 + H_{\text{int}},$$

$$H_0 = -\frac{1}{2} \mathbf{I}_\mu \cdot \mathbf{I}_\mu + \frac{1}{2} \frac{\partial\mathbf{b}_\mu}{\partial x_j} \cdot \frac{\partial\mathbf{b}_\mu}{\partial x_j} + \bar{\psi}(\gamma_j \partial_j + m)\psi, \quad (21)$$

$$H_{\text{int}} = 2\epsilon(\mathbf{b}_i \times \mathbf{b}_4) \cdot \mathbf{I}_i - 2\epsilon(\mathbf{b}_\mu \times \mathbf{b}_j) \cdot (\partial\mathbf{b}_\mu/\partial x_j) + \epsilon^2(\mathbf{b}_i \times \mathbf{b}_j)^2 - \mathbf{J}_\mu \cdot \mathbf{b}_\mu.$$

The quantized form of the supplementary condition is the same as in quantum electrodynamics.

<sup>9</sup> W. Heisenberg and W. Pauli, Z. Physik 56, 1 (1929).

PROPERTIES OF THE **b** QUANTA

The quanta of the **b** field clearly have spin unity and isotopic spin unity. We know their electric charge too because all the interactions that we proposed must satisfy the law of conservation of electric charge, which is exact. The two states of the nucleon, namely proton and neutron, differ by charge unity. Since they can transform into each other through the emission or absorption of a **b** quantum, the latter must have three charge states with charges  $\pm e$  and 0. Any measurement of electric charges of course involves the electromagnetic field, which necessarily introduces a preferential direction in isotopic space at all space-time points. Choosing the isotopic gauge such that this preferential direction is along the  $z$  axis in isotopic space, one sees that for the nucleons

$$Q = \text{electric charge} = e(\frac{1}{2} + \epsilon^{-1} T^z),$$

and for the **b** quanta

$$Q = (e/\epsilon) T^z.$$

The interaction (7) then fixes the electric charge up to an additive constant for all fields with any isotopic spin:

$$Q = e(\epsilon^{-1} T^z + R). \quad (22)$$

The constants  $R$  for two charge conjugate fields must be equal but have opposite signs.<sup>10</sup>

FIG. 1. Elementary vertices for **b** fields and nucleon fields. Dotted lines refer to **b** field, solid lines with arrow refer to nucleon field.



We next come to the question of the mass of the **b** quantum, to which we do not have a satisfactory answer. One may argue that without a nucleon field the Lagrangian would contain no quantity of the dimension of a mass, and that therefore the mass of the **b** quantum in such a case is zero. This argument is however subject to the criticism that, like all field theories, the **b** field is beset with divergences, and dimensional arguments are not satisfactory.

One may of course try to apply to the **b** field the methods for handling infinities developed for quantum electrodynamics. Dyson's approach<sup>11</sup> is best suited for the present case. One first transforms into the interaction representation in which the state vector  $\Psi$

<sup>10</sup> See M. Gell-Mann, Phys. Rev. **92**, 833 (1953).

<sup>11</sup> F. J. Dyson, Phys. Rev. **75**, 486, 1736 (1949).

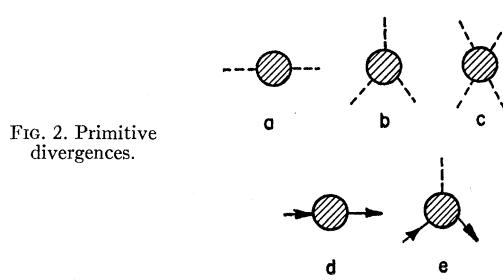


FIG. 2. Primitive divergences.

satisfies

$$i\partial\Psi/\partial t = H_{\text{int}}\Psi,$$

where  $H_{\text{int}}$  was defined in Eq. (21). The matrix elements of the scattering matrix are then formulated in terms of contributions from Feynman diagrams. These diagrams have three elementary types of vertices illustrated in Fig. 1, instead of only one type as in quantum electrodynamics. The "primitive divergences" are still finite in number and are listed in Fig. 2. Of these, the one labeled *a* is the one that effects the propagation function of the **b** quantum, and whose singularity determines the mass of the **b** quantum. In electrodynamics, by the requirement of electric charge conservation,<sup>12</sup> it is argued that the mass of the photon vanishes. Corresponding arguments in the **b** field case do not exist<sup>13</sup> even though the conservation of isotopic spin still holds. We have therefore not been able to conclude anything about the mass of the **b** quantum.

A conclusion about the mass of the **b** quantum is of course very important in deciding whether the proposal of the existence of the **b** field is consistent with experimental information. For example, it is inconsistent with present experiments to have their mass less than that of the pions, because among other reasons they would then be created abundantly at high energies and the charged ones should live long enough to be seen. If they have a mass greater than that of the pions, on the other hand, they would have a short lifetime (say, less than  $10^{-20}$  sec) for decay into pions and photons and would so far have escaped detection.

<sup>12</sup> J. Schwinger, Phys. Rev. **76**, 790 (1949).

<sup>13</sup> In electrodynamics one can formally prove that  $G_{\mu\nu}k_\nu = 0$ , where  $G_{\mu\nu}$  is defined by Schwinger's Eq. (A12). ( $G_{\mu\nu}A_\nu$  is the current generated through virtual processes by the arbitrary external field  $A_\nu$ ) No corresponding proof has been found for the present case. This is due to the fact that in electrodynamics the conservation of charge is a consequence of the equation of motion of the electron field alone, quite independently of the electromagnetic field itself. In the present case the **b** field carries an isotopic spin and destroys such general conservation laws.



# “Relative State” Formulation of Quantum Mechanics\*

Hugh Everett, III†

*Palmer Physical Laboratory, Princeton University, Princeton, New Jersey*

## 1. Introduction

The task of quantizing general relativity raises serious questions about the meaning of the present formulation and interpretation of quantum mechanics when applied to so fundamental a structure as the space-time geometry itself. This paper seeks to clarify the foundations of quantum mechanics. It presents a reformulation of quantum theory in a form believed suitable for application to general relativity.

The aim is not to deny or contradict the conventional formulation of quantum theory, which has demonstrated its usefulness in an overwhelming variety of problems, but rather to supply a new, more general and complete formulation, from which the conventional interpretation can be *deduced*.

The relationship of this new formulation to the older formulation is therefore that of a metatheory to a theory, that is, it is an underlying theory in which the nature and consistency, as well as the realm of applicability, of the older theory can be investigated and clarified.

---

\*Thesis submitted to Princeton University March 1, 1957 in partial fulfillment of the requirements for the Ph.D. degree. An earlier draft dated January, 1956 was circulated to several physicists whose comments were helpful. Professor Niels Bohr, Dr. H. J. Groenewald, Dr. Aage Peterson, Dr. A. Stern, and Professor L. Rosenfeld are free of any responsibility, but they are warmly thanked for the useful objections that they raised. Most particular thanks are due to Professor John A. Wheeler for his continued guidance and encouragement. Appreciation is also expressed to the National Science Foundation for fellowship support.

†Present address: Weapons Systems Evaluation Group, The Pentagon, Washington, D. C.

The new theory is not based on any radical departure from the conventional one. The special postulates in the old theory which deal with observation are omitted in the new theory. The altered theory thereby acquires a new character. It has to be analyzed in and for itself before any identification becomes possible between the quantities of the theory and the properties of the world of experience. The identification, when made, leads back to the omitted postulates of the conventional theory that deal with observation, but in a manner which clarifies their role and logical position.

We begin with a brief discussion of the conventional formulation, and some of the reasons which motivate one to seek a modification.

## 2. Realm of Applicability of the Conventional or “External Observation” Formulation of Quantum Mechanics

We take the conventional or “external observation” formulation of quantum mechanics to be essentially the following<sup>1</sup>: A physical system is completely described by a state function  $\psi$ , which is an element of a Hilbert space, and which furthermore gives information only to the extent of specifying the probabilities of the results of various observations which can be made *on* the system *by* external observers. There are two fundamentally different ways in which the state function can change:

*Process 1*: The discontinuous change brought about by the observation of a quantity with eigenstates  $\phi_1, \phi_2, \dots$ , in which the state  $\psi$  will be changed to the state  $\phi_j$  with probability  $|(\psi, \phi_j)|^2$ .

*Process 2*: The continuous, deterministic change of state of an isolated system with time according to a wave equation  $\partial\psi/\partial t = A\psi$ , where  $A$  is a linear operator.

This formulation describes a wealth of experience. No experimental evidence is known which contradicts it.

---

<sup>1</sup>We use the terminology and notation of J. von Neumann, *Mathematical Foundations of Quantum Mechanics*, translated by R. T. Beyer (Princeton University Press, Princeton, 1955).

Not all conceivable situations fit the framework of this mathematical formulation. Consider for example an isolated system consisting of an observer or measuring apparatus, plus an object system. Can the change with time of the state of the *total* system be described by Process 2? If so, then it would appear that no discontinuous probabilistic process like Process 1 can take place. If not, we are forced to admit that systems which contain observers are not subject to the same kind of quantum-mechanical description as we admit for all other physical systems. The question cannot be ruled out as lying in the domain of psychology. Much of the discussion of "observers" in quantum mechanics has to do with photoelectric cells, photographic plates, and similar devices where a mechanistic attitude can hardly be contested. For the following one can *limit himself to this class of problems*, if he is unwilling to consider observers in the more familiar sense on the same mechanistic level of analysis.

What mixture of Process 1 and 2 of the conventional formulation is to be applied to the case where only an approximate measure is effected; that is, where an apparatus or observer interacts only weakly and for a limited time with an object system? In this case of an approximate measurement a proper theory must specify (1) the new state of the object system that corresponds to any particular reading of the apparatus and (2) the probability with which this reading will occur. von Neumann showed how to treat a special class of approximate measurements by the method of projection operators.<sup>2</sup> However, a general treatment of all approximate measurements by the method of projections operators can be shown (Sec. 4) to be impossible.

How is one to apply the conventional formulation of quantum mechanics to the space-time geometry itself? The issue becomes especially acute in the case of a closed universe.<sup>3</sup> There is no place to stand outside the system to observe it. There is nothing outside it to produce transitions from one state to another. Even the familiar concept of a proper state of the energy is completely inapplicable. In the derivation of the law of conservation of energy, one defines the total energy by way of an integral extended over a surface large enough to include all parts of the system and their interactions.<sup>4</sup> But in a closed space, when a surface is made to include more and more of

---

<sup>2</sup>Reference 1, Chap. 4, Sec. 4.

<sup>3</sup>See A. Einstein, *The Meaning of Relativity* (Princeton University Press, Princeton, 1950), third edition, p. 107.

<sup>4</sup>L. Landau and E. Lifshitz, *The Classical Theory of Fields*, translated by M. Hamermesh (Addison-Wesley Press, Cambridge, 1951), p. 343.

the volume, it ultimately disappears into nothingness. Attempts to define the total energy for a closed space collapse to the vacuous statement, zero equals zero.

How are a quantum description of a closed universe, of approximate measurements, and of a system that contains an observer to be made? These three questions have one feature in common, that they all inquire about the *quantum mechanics* that is *internal to an isolated system*.

No way is evident to apply the conventional formulation of quantum mechanics to a system that is not subject to *external* observation. The whole interpretive scheme of that formalism rests upon the notion of external observation. The probabilities of the various possible outcomes of the observation are prescribed exclusively by Process 1. Without that part of the formalism there is no means whatever to ascribe a physical interpretation to the conventional machinery. But Process 1 is out of the question for systems not subject to external observation.<sup>5</sup>

### 3. Quantum Mechanics Internal to an Isolated System

This paper proposes to regard pure wave mechanics (Process 2 only) as a complete theory. It postulates that a wave function that obeys a linear wave equation everywhere and at all times supplies a complete mathematical model for every isolated physical system without exception. It further postulates that every system that is subject to external observation can be regarded as part of a larger isolated system.

The wave function is taken as the basic physical entity with *no a priori interpretation*. Interpretation only comes *after* an investigation of the logical structure of the theory. Here as always the theory itself sets the framework for its interpretation.

For any interpretation it is necessary to put the mathematical model of the theory into correspondence with experience. For this purpose it is necessary to formulate abstract models for observers that can be treated within the theory itself as physical systems, to consider isolated systems containing such model observers in interaction with other subsystems, to

---

<sup>5</sup>See in particular the discussion of this point by N. Bohr and L. Rosenfeld, Kgl. Danske Videnskab. Selskab, Mat.-fys. Medd. **12**, No. 8 (1933).

deduce the changes that occur in an observer as a consequence of interaction with the surrounding subsystems, and to interpret the changes in the familiar language of experience.

Section 4 investigates representations of the state of a composite system in terms of states of constituent subsystems. The mathematics leads one to recognize the concept of the *relativity of states*, in the following sense: a constituent subsystem cannot be said to be in any single well-defined state, independently of the remainder of the composite system. To any arbitrarily chosen state for one subsystem there will correspond a unique *relative state* for the remainder of the composite system. This relative state will usually depend upon the choice of state for the first subsystem. Thus the state of one subsystem does not have an independent existence, but is fixed only by the state of the remaining subsystem. In other words, the states occupied by the subsystems are not independent, but *correlated*. Such correlations between systems arise whenever systems interact. In the present formulation all measurements and observation processes are to be regarded simply as interactions between the physical systems involved—interactions which produce strong correlations. A simple model for a measurement, due to von Neumann, is analyzed from this viewpoint.

Section 5 gives an abstract treatment of the problem of observation. This uses only the superposition principle, and general rules by which composite system states are formed of subsystem states, in order that the results shall have the greatest generality and be applicable to any form of quantum theory for which these principles hold. Deductions are drawn about the state of the observer relative to the state of the object system. It is found that experiences of the observer (magnetic tape memory, counter system, etc.) are in full accord with predictions of the conventional “external observer” formulation of quantum mechanics, based on Process 1.

Section 6 recapitulates the “relative state” formulation of quantum mechanics.

## 4. Concept of Relative State

We now investigate some consequences of the wave mechanical formalism of composite systems. If a composite system  $S$ , is composed of two subsystems  $S_1$  and  $S_2$ , with associated Hilbert spaces  $H_1$  and  $H_2$ , then, according to the usual formalism of composite systems, the Hilbert space for  $S$  is taken

to be the *tensor product* of  $H_1$  and  $H_2$  (written  $H = H_1 \otimes H_2$ ). This has the consequence that if the sets  $\{\xi_i^{S_1}\}$  and  $\{\eta_j^{S_2}\}$  are complete orthonormal sets of states for  $S_1$  and  $S_2$ , respectively, then the general state of  $S$  can be written as a superposition:

$$\psi^S = \sum_{i,j} a_{ij} \xi_i^{S_1} \eta_j^{S_2}. \quad (1)$$

From (3.1) [sic] although  $S$  is in a definite state  $\psi^S$ , the subsystems  $S_1$  and  $S_2$  do not possess anything like definite states independently of one another (except in the special case where all but one of the  $a_{ij}$  are zero).

We can, however, for any choice of a state in one subsystem, uniquely assign a corresponding *relative state* in the other subsystem. For example, if we choose  $\xi_k$  as the state for  $S_1$ , while the composite system  $S$  is in the state  $\psi^S$  given by (3.1) [sic], then the corresponding *relative state* in  $S_2$ ,  $\psi(S_2; \text{rel}\xi_k, S_1)$ , will be:

$$\psi(S_2; \text{rel}\xi_k, S_1) = N_k \sum_j a_{kj} \eta_j^{S_2} \quad (2)$$

where  $N_k$  is a normalization constant. This relative state for  $\xi_k$  is independent of the choice of basis  $\{\xi_i\}$  ( $i \neq k$ ) for the orthogonal complement of  $\xi_k$ , and is hence determined uniquely by  $\xi_k$  alone. To find the relative state in  $S_2$  for an arbitrary state of  $S_1$  therefore, one simply carries out the above procedure using any pair of bases for  $S_1$  and  $S_2$  which contains the desired state as one element of the basis for  $S_1$ . To find states in  $S_1$  relative to states in  $S_2$ , interchange  $S_1$  and  $S_2$  in the procedure.

In the conventional or “external observation” formulation, the relative state in  $S_2$ ,  $\psi(S_2; \text{rel}\phi, S_1)$ , for a state  $\phi^{S_1}$  in  $S_1$ , gives the conditional probability distributions for the results of all measurements in  $S_2$ , given that  $S_1$  has been measured and found to be in state  $\phi^{S_1}$ —i.e., that  $\phi^{S_1}$  is the eigenfunction of the measurement in  $S_1$  corresponding to the observed eigenvalue.

For any choice of basis in  $S_1$ ,  $\{\xi_i\}$ , it is always possible to represent the state of  $S$ , (1), as a *single* superposition of pairs of states, each consisting of a state from the basis  $\{\xi_i\}$  in  $S_1$  and its relative state in  $S_2$ . Thus, from (2), (1) can be written in the form:

$$\psi^S = \sum_i \frac{1}{N_i} \xi_i^{S_1} \psi(S_2; \text{rel}\xi_i, S_1). \quad (3)$$

This is an important representation used frequently

*Summarizing: There does not, in general, exist anything like a single state for one subsystem of a composite system. Subsystems do not possess states that are independent of the states of the remainder of the system, so that the subsystem states are generally correlated with one another. One can arbitrarily choose a state for one subsystem, and be led to the relative state for the remainder. Thus we are faced with a fundamental relativity of states, which is implied by the formalism of composite systems. It is meaningless to ask the absolute state of a subsystem—one can ask the state relative to a given state of the remainder of the subsystem.*

At this point we consider a simple example, due to von Neumann, which serves as a model of a measurement process. Discussion of this example prepares the ground for the analysis of “observation.” We start with a system of only one coordinate,  $q$  (such as position of a particle), and an apparatus of one coordinate  $r$  (for example the position of a meter needle). Further suppose that they are initially independent, so that the combined wave function is  $\psi_0^{S+A} = \phi(q)\eta(r)$  where  $\phi(q)$  is the initial system wave function, and  $\eta(r)$  is the initial apparatus function. The Hamiltonian is such that the two systems do not interact except during the interval  $t = 0$  to  $t = T$ , during which time the total Hamiltonian consists only of a simple interaction,

$$H_I = -i\hbar q(\partial/\partial r). \quad (4)$$

Then the state

$$\psi_t^{S+A}(q, r) = \phi(q)\eta(r - qt) \quad (5)$$

is a solution of the Schrödinger equation,

$$i\hbar(\partial\psi_t^{S+A}/\partial t) = H_I\psi_t^{S+A} \quad (6)$$

for the specified initial conditions at time  $t = 0$ .

From (5) at time  $t = T$  (at which time interaction stops) there is no longer any definite independent apparatus state, nor any independent system state. The apparatus therefore does not indicate any definite object-system value, and nothing like process 1 has occurred.

Nevertheless, we *can* look upon the total wave function (5) as a *superposition* of pairs of subsystem states, each element of which has a definite  $q$  value and a correspondingly displaced apparatus state. Thus after the interaction the state (5) has the form:

$$\psi_T^{S+A} = \int \phi(q')\delta(q - q')\eta(r - qT)dq', \quad (7)$$

which is a superposition of states  $\psi_{q'} = \delta(q - q')\eta(r - qT)$ . Each of these elements,  $\psi_{q'}$ , of the superposition describes a state in which the system has the definite value  $q = q'$ , and in which the apparatus has a state that is displaced from its original state by the amount  $q'T$ . These elements  $\psi_{q'}$  are then superposed with coefficients  $\phi(q')$  to form the total state (7).

Conversely, if we transform to the representation where the *apparatus* coordinate is definite, we write (5) as

$$\psi_T^{S+A} = \int (1/N_{r'}) \xi^{r'}(q) \delta(r - r') dr',$$

where

$$\xi^{r'}(q) = N_{r'} \phi(q) \eta(r' - qT) \quad (8)$$

and

$$(1/N_{r'})^2 = \int \phi^*(q) \phi(q) \eta^*(r' - qT) \eta(r' - qT) dq.$$

Then the  $\xi^{r'}(q)$  are the relative system state functions<sup>6</sup> for the apparatus states  $\delta(r - r')$  of definite value  $r = r'$ .

If  $T$  is sufficiently large, or  $\eta(r)$  sufficiently sharp (near  $\delta(r)$ ), then  $\xi^{r'}(q)$  is nearly  $\delta(q - r'/T)$  and the relative system states  $\xi^{r'}(q)$  are nearly eigenstates for the values  $q = r'/T$ .

We have seen that (8) is a superposition of states  $\psi_{r'}$ , *for each of which* the apparatus has recorded a definite value  $r'$ , and the system is left in approximately the eigenstate of the measurement corresponding to  $q = r'/T$ . The discontinuous “jump” into an eigenstate is thus only a relative proposition, dependent upon the mode of decomposition of the total wave function into the superposition, and relative to a particularly chosen apparatus-coordinate value. So far as the complete theory is concerned all elements of the superposition exist simultaneously, and the entire process is quite continuous.

von Neumann’s example is only a special case of a more general situation. Consider any measuring apparatus interacting with any object system. As

---

<sup>6</sup>This example provides a model of an approximate measurement. However, the relative system states after the interaction  $\xi^{r'}(q)$  cannot ordinarily be generated from the original system state  $\phi$  by the application of *any* projection operator,  $E$ . Proof: Suppose on the contrary that  $\xi^{r'}(q) = NE\phi(q) = N'\phi(q)\eta(r' - qt)$ , where  $N, N'$  are normalization constants. Then

$$E(NE\phi(q)) = NE^2\phi(q) = N''\phi(q)\eta^2(r' - qt)$$

and  $E^2\phi(q) = (N''/N)\phi(q)\eta^2(r' - qt)$ . But the condition  $E^2 = E$  which is necessary for  $E$  to be a projection implies that  $N'/N''\eta(q) = \eta^2(q)$  which is generally false.

a result of the interaction the state of the measuring apparatus is no longer capable of independent definition. It can be defined only *relative* to the state of the object system. In other words, there exists only a correlation between the two states of the two systems. It seems as if nothing can ever be settled by such a measurement.

This indefinite behavior seems to be quite at variance with our observations, since physical objects always appear to us to have definite positions. Can we reconcile this feature wave mechanical theory built purely on Process 2 with experience, or must the theory be abandoned as untenable? In order to answer this question we consider the problem of observation itself within the framework of the theory.

## 5. Observation

We have the task of making deductions about the appearance of phenomena to observers which are considered as purely physical systems and are treated within the theory. To accomplish this it is necessary to identify some present properties of such an observer with features of the past experience of the observer. Thus, in order to say that an observer 0 has observed the event  $\alpha$ , it is necessary that the state of 0 has become changed from its former state to a new state which is dependent upon  $\alpha$ .

It will suffice for our purposes to consider the observers to possess memories (i.e., parts of a relatively permanent nature whose states are in correspondence with past experience of the observers). In order to make deductions about the past experience of an observer it is sufficient to deduce the present contents of the memory as it appears within the mathematical model.

As models for observers we can, if we wish, consider automatically functioning machines, possessing sensory apparatus and coupled to recording devices capable of registering past sensory data and machine configurations. We can further suppose that the machine is so constructed that its present actions shall be determined not only by its present sensory data, but by the contents of its memory as well. Such a machine will then be capable of performing a sequence of observations (measurements), and furthermore of deciding upon its future experiments on the basis of past results. If we consider that current sensory data, as well as machine configuration, is immediately recorded in the memory, then the actions of the machine at a given instant can be regarded as a function of the memory contents only, and all

relevant [*sic*] experience of the machine is contained in the memory 

For such machines we are justified in using such phrases as “the machine has perceived  $A$ ” or “the machine is aware of  $A$ ” if the occurrence of  $A$  is represented in the memory, since the future behavior of the machine will be based upon the occurrence of  $A$ . In fact, all of the customary language of subjective experience is quite applicable to such machines, and forms the most natural and useful mode of expression when dealing with their behavior, as is well known to individuals who work with complex automata.

When dealing with a system representing an observer quantum mechanically we ascribe a state function,  $\psi^0$ , to it. When the state  $\psi^0$  describes an observer whose memory contains representations of the events  $A, B, \dots, C$  we denote this fact by appending the memory sequence in brackets as a subscript, writing:

$$\psi_{[A, B, \dots, C]}^0. \quad (9)$$

The symbols  $A, B, \dots, C$ , which we assume to be ordered time-wise, therefore stand for memory configurations which are in correspondence with the past experience of the observer. These configurations can be regarded as punches in a paper tape, impressions on a magnetic reel, configurations of a relay switching circuit, or even configurations of brain cells. We require only that they be capable of the interpretation “The observer has experienced the succession of events  $A, B, \dots, C$ . (We sometimes write dots in a memory sequence,  $\dots A, B, \dots, C$ , to indicate the possible presence  of previous memories which are irrelevant to the case being considered.)

The mathematical model seeks to treat the interaction of such observer systems with other physical systems (observations), within the framework of Process 2 wave mechanics, and to deduce the resulting memory configurations, which  then to be interpreted as records of the past experiences of the observers.

We begin by defining what constitutes a “good” observation. A good observation of a quantity  $A$ , with eigenfunctions  $\phi_i$ , for a system  $S$ , by an observer whose initial states is  $\psi^0$ , consists of an interaction which, in a specified period of time, transforms each (total) state

$$\psi^{S+0} = \phi_i \psi_{[\dots]}^0 \quad (10)$$

into a new state

$$\psi^{S+0'} = \phi_i \psi_{[\dots\alpha_i]}^0 \quad (11)$$

where  $\alpha_i$  characterizes<sup>7</sup> the state  $\phi_i$ . (The symbol,  $\alpha_i$ , might stand for a recording of the eigenvalue, for example.) That is, we require that the system state, *if it is an eigenstate*, shall be unchanged, and (2) that the observer state shall change so as to describe an observer that is “aware” of which eigenfunction it is; that is, some property is recorded in the memory of the observer which characterizes  $\phi_i$ , such as the eigenvalue. The requirement that the eigenstates for the system be unchanged is necessary if the observation is to be significant (repeatable), and the requirement that the observer state change in a manner which is different for each eigenfunction is necessary if we are to be able to call the interaction an observation at all. How closely a general interaction satisfies the definition of a good observation depends upon (1) the way in which the interaction depends upon the dynamical variables of the observer system—including memory variables—and upon the dynamical variables of the object system and (2) the initial state of the observer system. Given (1) and (2), one can for example solve the wave equation, deduce the state of the composite system after the end of the interaction, and check whether an object system that was originally in an eigenstate is left in an eigenstate, as demanded by the repeatability postulate. This postulate is satisfied, for example, by the model of von Neumann that has already been discussed.

From the definition of a good observation we first deduce the result of an observation upon a system which is *not* in an eigenstate of the observation. We know from our definition that the interaction transforms states  $\phi_i \psi^0_{[...]} \blacksquare$  into states  $\phi_i \psi^0_{[... \alpha_i]} \blacksquare$ . Consequently these solutions of the wave equation can be superposed to give the final state for the case of an arbitrary initial system state. Thus if the initial system state is not an eigenstate, but a general state  $\sum_i a_i \phi_i$ , the final total state will have the form:

$$\psi^{S+0'} = \sum_i a_i \phi_i \psi^0_{[... \alpha_i]} \blacksquare \quad (12)$$

This superposition principle continues to apply in the presence of further systems which do not interact during the measurement. Thus, if systems  $S_1, S_2, \dots, S_n$  are present as well as 0, with original states  $\psi^{S_1}, \psi^{S_2}, \dots, \psi^{S_n}$ , and the only interaction during the time of measurement takes place between

---

<sup>7</sup>It should be understood that  $\psi^0_{[... \alpha_i]}$  is a *different* state for each  $i$ . A more precise notation would write  $\psi^0_{i[... \alpha_i]}$ , but no confusion can arise if we simply let the  $\psi^0_i$  be indexed only by the index of the memory configuration symbol.

$S_1$  and 0, the measurement will transform the initial total state:

$$\psi^{S_1+S_2+\dots+S_n+0} = \psi^{S_1}\psi^{S_2}\dots\psi^{S_n}\psi_{[\dots]}^0 \quad (13)$$

into the final state:

$$\psi'^{S_1+S_2+\dots+S_n+0} = \sum_i a_i \phi_i^{S_1} \psi^{S_2} \dots \psi^{S_n} \psi_{[\dots\alpha_i]}^0 \quad (14)$$

where  $a_i = (\phi_i^{S_1}, \psi^{S_1})$  and  $\phi_i^{S_1}$  are the eigenfunctions of the observation 

Thus we arrive at the general rule for the transformation of total state functions which describe systems within which observation processes occur:

*Rule 1:* The observation of a quantity  $A$ , with eigenfunctions  $\phi_i^{S_1}$ , in a system  $S_1$  by the observer 0, transforms the total state according to:

$$\psi^{S_1}\psi^{S_2}\dots\psi^{S_n}\psi_{[\dots]}^0 \rightarrow \sum_i a_i \phi_i^{S_1} \psi^{S_2} \dots \psi^{S_n} \psi_{[\dots\alpha_i]}^0 \quad (15)$$

where

$$a_i = (\phi_i^{S_1}, \psi^{S_1}).$$

If we next consider a *second* observation to be made, where our total state is now a superposition, we can apply Rule 1 separately to each element of the superposition, since each element separately obeys the wave equation and behaves independently of the remaining elements, and then superpose the results to obtain the final solution. We formulate this as:

*Rule 2:* Rule 1 may be applied separately to each element of a superposition of total system states, the results being superposed to obtain the final total state. Thus, a determination of  $B$ , with eigenfunctions  $\eta_j^{S_2}$ , on  $S_2$  by the observer 0 transforms the total state

$$\sum_i a_i \phi_i^{S_1} \psi^{S_2} \dots \psi^{S_n} \psi_{[\dots\alpha_i]}^0 \quad (16)$$

into the state

$$\sum_{i,j} a_i b_j \phi_i^{S_1} \eta_j^{S_2} \psi^{S_3} \dots \psi^{S_n} \psi_{[\dots\alpha_i, \beta_j]}^0 \quad (17)$$

where  $b_j = (\eta_j^{S_2}, \psi^{S_2})$ , which follows from the application of Rule 1 to each element  $\phi_i^{S_1} \psi^{S_2} \dots \psi^{S_n} \psi_{[\dots\alpha_i]}^0$ , and then superposing the results with the coefficients  $a_i$ . 

These two rules, which follow directly from the superposition principle, give a convenient method for determining final total states for any number of observation processes in any combinations. We now seek the *interpretation* of such final total states.

Let us consider the simple case of a single observation of a quantity  $A$ , with eigenfunctions  $\phi_i$ , in the system  $S$  with initial state  $\psi^S$ , by an observer 0 whose initial state is  $\psi_{[...]}^0$ . The final result is, as we have seen, the superposition

$$\psi'^{S+0} = \sum_i a_i \phi_i \psi_{[\dots\alpha_i]}^0. \quad (18)$$

There is no longer any independent system state or observer state, although the two have become correlated in a one-one manner. However, in each element of the superposition,  $\phi_i \psi_{[\dots\alpha_i]}^0$ , the object-system state is a particular eigenstate of the observation, and furthermore the observer-system state describes the observer as definitely perceiving that particular system state. This correlation is what allows one to maintain the interpretation that a measurement has been performed.

We now consider a situation where the observer system comes into interaction with the object system for a second time. According to Rule 2 we arrive at the total state after the second observation:

$$\psi''^{S+0} = \sum_i a_i \phi_i \psi_{[\dots\alpha_i, \alpha_i]}^0. \quad (19)$$

Again, each element  $\phi_i \psi_{[\dots\alpha_i, \alpha_i]}^0$  describes a system eigenstate, but this time also describes the observer as having obtained the *same result* for each of the two observations. Thus for every separate state of the observer in the final superposition the result of the observation was repeatable, even though different for different states. The repeatability is a consequence of the fact that after an observation the *relative* system state for a particular observer state is the corresponding eigenstate.

Consider now a different situation. An observer-system 0, with initial state  $\psi_{[...]}^0$ , measures the *same* quantity  $A$  in a number of separate, identical, systems which are initially in the same state,  $\psi^{S_1} = \psi^{S_2} = \dots = \psi^{S_n} = \sum_i a_i \phi_i$  (where the  $\phi_i$  are, as usual, eigenfunctions of  $A$ ). The initial total state function is then

$$\psi_0^{S_1 + S_2 + \dots + S_n + 0} = \psi^{S_1} \psi^{S_2} \dots \psi^{S_n} \psi_{[...]}^0. \quad (20)$$

We assume that the measurements are performed on the systems in the order  $S_1, S_2, \dots, S_n$ . Then the total state after the first measurement is by Rule 1,

$$\psi_1^{S_1+S_2+\dots+S_n+0} = \sum_i a_i \phi_i^{S_1} \psi^{S_2} \dots \psi^{S_n} \psi_{[\dots \alpha_i^1]}^0 \quad (21)$$

(where  $\alpha_i^1$  refers to the first system,  $S_1$ ).

After the second measurement it is, by Rule 2,

$$\psi_2^{S_1+S_2+\dots+S_n+0} = \sum_{i,j} a_i a_j \phi_i^{S_1} \phi_j^{S_2} \psi^{S_3} \dots \psi^{S_n} \psi_{[\dots \alpha_i^1, \alpha_j^2]}^0 \quad (22)$$

and in general, after  $r$  measurements have taken place ( $r \leq n$ ), Rule 2 gives the result:

$$\psi_r = \sum_{i,j,\dots,k} a_i a_j \dots a_k \phi_i^{S_1} \phi_j^{S_2} \dots \phi_k^{S_r} \psi^{S_{r+1}} \dots \psi^{S_n} \psi_{[\dots \alpha_i^1, \alpha_j^2, \dots, \alpha_k^r]}^0 \quad (23)$$

We can give this state,  $\psi_r$ , the following interpretation. It consists of a superposition of states:

$$\psi'_{ij\dots k} = \phi_i^{S_1} \phi_j^{S_2} \dots \phi_k^{S_r} \times \psi^{S_{r+1}} \dots \psi^{S_n} \psi_{[\alpha_i^1, \alpha_j^2, \dots, \alpha_k^r]}^0 \quad (24)$$

each of which describes the observer with a definite memory sequence  $[\alpha_i^1, \alpha_j^2, \dots, \alpha_k^r]$ . Relative to him the (observed) system states are the corresponding eigenfunctions  $\phi_i^{S_1}, \phi_j^{S_2}, \dots, \phi_k^{S_r}$ , the remaining systems,  $S_{r+1}, \dots, S_n$ , being unaltered.

A typical element  $\psi'_{ij\dots k}$  of the final superposition describes a state of affairs wherein the observer has perceived an apparently random sequence of definite results for the observations  Furthermore the object systems have been left in the corresponding eigenstates of the observation. At this stage suppose that a redetermination of an earlier system observation ( $S_l$ ) takes place. Then it follows that every element of the resulting final superposition will describe the observer with a memory configuration of the form  $[\alpha_i^1, \dots, \alpha_j^l, \dots, \alpha_k^r, \alpha_j^l]$  in which the earlier memory coincides with the later—i.e., the memory states are *correlated*. It will thus *appear* to the observer, as described by a typical element of the superposition, that each initial observation on a system caused the system to “jump” into an eigenstate in a random fashion and thereafter remain there for subsequent measurements

on the same system<sup>‡</sup>—therefore—disregarding for the moment quantitative questions of relative frequencies—the probabilistic assertions of Process 1 appear to be valid to the observer described by a typical element of the final superposition<sup>‡</sup>

We thus arrive at the following picture: Throughout all of a sequence of observation processes there is only one physical system representing the observer, yet there is no single unique *state* of the observer (which follows from the representations of interacting systems)<sup>‡</sup> Nevertheless, there is a representation in terms of a *superposition*, each element of which contains a definite observer state and a corresponding system state. Thus with each succeeding observation (or interaction), the observer state “branches” into a number of different states<sup>‡</sup> Each branch represents a different outcome of the measurement and the *corresponding* eigenstate for the object-system state. All branches exist simultaneously in the superposition after any given sequence of observations.<sup>‡</sup>

The “trajectory” of the memory configuration of an observer performing a sequence of measurements is thus not a linear sequence of memory configurations, but a branching tree, with all possible outcomes existing simultaneously in a final superposition with various coefficients in the mathematical

---

<sup>‡</sup>*Note added in proof.*—In reply to a preprint of this article some correspondents have raised the question of the “transition from possible to actual,” arguing that in “reality” there is—as our experience testifies—no such splitting of observer states, so that only one branch can ever actually exist. Since this point may occur to other readers the following is offered in explanation.

The whole issue of the transition from “possible” to “actual” is taken care of in the theory in a very simple way—there is no such transition, nor is such a transition necessary for the theory to be in accord with our experience. From the viewpoint of the theory *all* elements of a superposition (all “branches”) are “actual,” none any more “real” than the rest<sup>‡</sup> It is unnecessary to suppose that all but one are somehow destroyed, since all the separate elements of a superposition individually obey the wave equation with complete indifference to the presence or absence (“actuality” or not) of any other elements. This total lack of effect of one branch on another also implies that no observer will ever be aware of any “splitting” process<sup>‡</sup>

Arguments that the world picture presented by this theory is contradicted by experience, because we are unaware of any branching process, are like the criticism of the Copernican theory that the mobility of the earth as a real physical fact is incompatible with the common sense interpretation of nature because we feel no such motion. In both cases the argument fails when it is shown that the theory itself predicts that our experience will be what it in fact is.<sup>‡</sup> (In the Copernican case the addition of Newtonian physics was required to be able to show that the earth’s inhabitants would be unaware of any motion of the earth.)

model. In any familiar memory device the branching does not continue indefinitely, but must stop at a point limited by the capacity of the memory.

In order to establish quantitative results, we must put some sort of measure (weighting) on the elements of a final superposition. This is necessary to be able to make assertions which hold for almost all of the observer states described by elements of a superposition. We wish to make quantitative statements about the relative frequencies of the different possible results of observation—which are recorded in the memory—for a typical observer state; but to accomplish this we must have a method for selecting a typical element from a superposition of orthogonal states.

We therefore seek a general scheme to assign a measure to the elements of a superposition of orthogonal states  $\sum_i a_i \phi_i$ . We require a positive function  $m$  of the complex coefficients of the elements of the superposition, so that  $m(a_i)$  shall be the measure assigned to the element  $\phi_i$ . In order that this general scheme be unambiguous we must first require that the states themselves always be normalized, so that we can distinguish the coefficients from the states. However, we can still only determine the *coefficients*, in distinction to the states, up to an arbitrary phase factor. In order to avoid ambiguities the function  $m$  must therefore be a function of the amplitudes of the coefficients alone,  $m(a_i) = m(|a_i|)$ .

We now impose an additivity requirement. We can regard a subset of the superposition, say  $\sum_{i=1}^n a_i \phi_i$ , as a single element  $\alpha \phi'$ :

$$\alpha \phi' = \sum_{i=1}^n a_i \phi_i. \quad (25)$$

We demand that the measure assigned to  $\phi'$  shall be the sum of the measures assigned to the  $\phi_i$  ( $i$  from 1 to  $n$ ):

$$m(\alpha) = \sum_{i=1}^n m(a_i). \quad (26)$$

Then we have already restricted the choice of  $m$  to the square amplitude alone; in other words, we have  $m(a_i) = a_i^* a_i$ , apart from a multiplicative constant.

To see this, note that the normality of  $\phi'$  requires that  $|\alpha| = (\sum a_i^* a_i)^{\frac{1}{2}}$ . From our remarks about the dependence of  $m$  upon the amplitude alone, we replace the  $a_i$  by their amplitudes  $u_i = |a_i|$ . Equation (26) then imposes the

requirement,

$$m(\alpha) = m\left(\sum a_i^* a_i\right)^{\frac{1}{2}} = m\left(\sum u_i^2\right)^{\frac{1}{2}} = \sum m(u_i) = \sum m(u_i^2)^{\frac{1}{2}}. \quad (27)$$

Defining a new function  $g(x)$

$$g(x) = m(\sqrt{x}) \quad (28)$$

we see that (27) requires that

$$g\left(\sum u_i^2\right) = \sum g(u_i^2). \quad (29)$$

Thus  $g$  is restricted to be linear and necessarily has the form:

$$g(x) = cx \quad (c \text{ constant}). \quad (30)$$

Therefore  $g(x^2) = cx^2 = m(\sqrt{x^2}) = m(x)$  and we have deduced that  $m$  is restricted to the form

$$m(a_i) = m(u_i) = cu_i^2 = ca_i^* a_i. \quad (31)$$

We have thus shown that the only choice of measure consistent with our additivity requirement is the square amplitude measure, apart from an arbitrary multiplicative constant which may be fixed, if desired, by normalization requirements. (The requirement that the total measure be unity implies that this constant is 1.)

The situation here is fully analogous to that of classical statistical mechanics, where one puts a measure on trajectories of systems in the phase space by placing a measure on the phase space itself, and then making assertions (such as ergodicity, quasi-ergodicity, etc.) which hold for “almost all” trajectories. This notion of “almost all” depends here also upon the choice of measure, which is in this case taken to be the Lebesgue measure on the phase space. He could contradict the statements of classical statistical mechanics by choosing a measure for which only the exceptional trajectories had nonzero measure. Nevertheless the choice of Lebesgue measure on the phase space can be justified by the fact that it is the only choice for which the “conservation of probability” holds, (Liouville’s theorem) and hence the only choice which makes possible any reasonable statistical deductions at all.

In our case, we wish to make statements about “trajectories” of observers. However, for us a trajectory is constantly branching (transforming from state

to superposition) with each successive measurement. To have a requirement analogous to the “conservation of probability” in the classical case, we demand that the measure assigned to a trajectory at one time shall equal the sum of the measures of its separate branches at a later time. This is precisely the additivity requirement which we imposed and which leads uniquely to the choice of square-amplitude measure. Our procedure is therefore quite as justified as that of classical statistical mechanics.

Having deduced that there is a unique measure which will satisfy our requirements, the square-amplitude measure, we continue our deduction. This measure then assigns to the  $i, j, \dots, k$ th element of the superposition (24),

$$\phi_i^{S_1} \phi_j^{S_2} \cdots \phi_k^{S_r} \psi^{S_{r+1}} \cdots \psi^{S_n} \psi_{[\alpha_i^1, \alpha_j^2, \dots, \alpha_k^r]}^0 \quad (32)$$

the measure (weight)

$$M_{ij\dots k} = (a_i a_j \cdots a_k)^* (a_i a_j \cdots a_k) \quad (33)$$

so that the observer state with memory configuration  $[\alpha_i^1, \alpha_j^2, \dots, \alpha_k^r]$  is assigned the measure  $a_i^* a_i a_j^* a_j \cdots a_k^* a_k = M_{ij\dots k}$ . We see immediately that this is a product measure, namely,

$$M_{ij\dots k} = M_i M_j \cdots M_k \quad (34)$$

where

$$M_l = a_l^* a_l$$

so that the measure assigned to a particular memory sequence  $[\alpha_i^1, \alpha_j^2, \dots, \alpha_k^r]$  is simply the product of the measures for the individual components of the memory sequence.

There is a direct correspondence of our measure structure to the probability theory of random sequences. If we regard the  $M_{ij\dots k}$  as probabilities for the sequences then the sequences are equivalent to the random sequences which are generated by ascribing to each term the *independent* probabilities  $M_l = a_l^* a_l$ . Now probability theory is equivalent to measure theory mathematically, so that we can make use of it, while keeping in mind that all results should be translated back to measure theoretic language.

Thus, in particular, if we consider the sequences to become longer and longer (more and more observations performed) each memory sequence of the final superposition will satisfy any given criterion for a randomly generated sequence, generated by the independent probabilities  $a_l^* a_l$ , except for a

set of total measure which tends toward zero as the number of observations becomes unlimited. Hence all averages of functions over *any* memory sequence, including the special case of frequencies, can be computed from the probabilities  $a_i^* a_i$ , except for a set of memory sequences of measure zero. We have therefore shown that the statistical assertions of Process 1 will appear to be valid to the observer, *in almost all* elements of the superposition (24), in the limit as the number of observations goes to infinity.

While we have so far considered only sequences of observations of the same quantity upon identical systems, the result is equally true for arbitrary sequences of observations, as may be verified by writing more general sequences of measurements, and applying Rules 1 and 2 in the same manner as presented here.

We can therefore summarize the situation when the sequence of observations is arbitrary, when these observations are made upon the same or different systems in any order, and when the number of observations of each quantity in each system is very large, with the following result:

Except for a set of memory sequences of measure nearly zero, the averages of any functions over a memory sequence can be calculated approximately by the use of the independent probabilities given by Process 1 for each initial observation, on a system, and by the use of the usual transition probabilities for succeeding observations upon the same system. In the limit, as the number of all types of observations goes to infinity the calculation is exact, and the exceptional set has measure zero.

This prescription for the calculation of averages over memory sequences by probabilities assigned to individual elements is precisely that of the conventional “external observation” theory (Process 1). Moreover, these predictions hold for almost all memory sequences. Therefore all predictions of the usual theory will appear to be valid to the observer in almost [sic] all observer states.

In particular, the uncertainty principle is never violated since the latest measurement upon a system supplies all possible information about the relative system state, so that there is no direct correlation between any earlier results of observation on the system, and the succeeding observation. Any observation of a quantity  $B$ , between two successive observations of quantity  $A$  (all on the same system) will destroy the one-one correspondence between the earlier and later memory states for the result of  $A$ . Thus for alternating

observations of different quantities there are fundamental limitations upon the correlations between memory states for the same observed quantity, these limitations expressing the content of the uncertainty principle.

As a final step one may investigate the consequences of allowing several observer systems to interact with (observe) the same object system, as well as to interact with one another (communicate). The latter interaction can be treated simply as an interaction which correlates parts of the memory configuration of one observer with another. When these observer systems are investigated, in the same manner as we have already presented in this section using Rules 1 and 2, one finds that in *all elements* of the final superposition:

1. When several observers have separately observed the same quantity in the object system and then communicated the results to one another they find that they are in agreement. This agreement persists even when an observer performs his observation *after* the result has been communicated to him by another observer who has performed the observation.

2. Let one observer perform an observation of a quantity  $A$  in the object system, then let a second perform an observation of a quantity  $B$  in this object system which does not commute with  $A$ , and finally let the first observer repeat his observation of  $A$ . Then the memory system of the first observer will *not* in general show the same result for both observations. The intervening observation by the other observer of the non-commuting quantity  $B$  prevents the possibility of any one to one correlation between the two observations of  $A$ .

3. Consider the case where the states of two object systems are correlated, but where the two systems do not interact. Let one observer perform a specified observation on the first system, then let another observer perform an observation on the second system, and finally let the first observer repeat his observation. Then it is found that the first observer always gets the same result both times, and the observation by the second observer has no effect whatsoever on the outcome of the first's observations. Fictitious paradoxes like that of Einstein, Podolsky, and Rosen<sup>8</sup> which are concerned with such correlated, noninteracting systems are easily investigated and clarified in the present scheme.

---

<sup>8</sup>Einstein, Podolsky, and Rosen, Phys. Rev. **47**, 777 (1935). For a thorough discussion of the physics of observation, see the chapter by N. Bohr in *Albert Einstein, Philosopher-Scientist* (The Library of Living Philosophers, Inc., Evanston, 1949).

Many further combinations of several observers and systems can be studied within the present framework. The results of the present “relative state” formalism agree with those of the conventional “external observation” formalism in all those cases where that familiar machinery is applicable

In conclusion, the continuous evolution of the state function of a composite system with time gives a complete mathematical model for processes that involve an idealized observer. When interaction occurs, the result of the evolution in time is a superposition of states, each element of which assigns a different state to the memory of the observer. Judged by the state of the memory in almost all of the observer states, the probabilistic conclusion [*sic*] of the usual “external observation” formulation of quantum theory are valid. In other words, pure Process 2 wave mechanics, without any initial probability assertions, leads to all the probability concepts of the familiar formalism.

## 6. Discussion

The theory based on pure wave mechanics is a conceptually simple, causal theory, which gives predictions in accord with experience. It constitutes a framework in which one can investigate in detail, mathematically, and in a logically consistent manner a number of sometimes puzzling subjects, such as the measuring process itself and the interrelationship of several observers. Objections have been raised in the past to the conventional or “external observation” formulation of quantum theory on the grounds that its probabilistic features are postulated in advance instead of being derived from the theory itself. We believe that the present “relative-state” formulation meets this objection, while retaining all of the content of the standard formulation.

While our theory ultimately justifies the use of the probabilistic interpretation as an aid to making practical predictions, it forms a broader frame in which to understand the consistency of that interpretation. In this respect it can be said to form a *metatheory* for the standard theory. It transcends the usual “external observation” formulation, however, in its ability to deal logically with questions of imperfect observation and approximate measurement.

The “relative state” formulation will apply to all forms of quantum mechanics which maintain the superposition principle. It may therefore prove a

fruitful framework for the quantization of general relativity. The formalism invites one to construct the formal theory first, and to supply the statistical interpretation later. This method should be particularly useful for interpreting quantized unified field theories where there is no question of ever isolating observers and object systems. They all are represented in a *single* structure, the field. Any interpretative rules can probably only be deduced in and through the theory itself.

Aside from any possible practical advantages of the theory, it remains a matter of intellectual interest that the statistical assertions of the usual interpretation do not have the status of independent hypotheses, but are deducible (in the present sense) from the pure wave mechanics that starts completely free of statistical postulates.

## Dynamical Model of Elementary Particles Based on an Analogy with Superconductivity. I\*

Y. NAMBU AND G. JONA-LASINIO†

*The Enrico Fermi Institute for Nuclear Studies and the Department of Physics, The University of Chicago, Chicago, Illinois*

(Received October 27, 1960)

It is suggested that the nucleon mass arises largely as a self-energy of some primary fermion field through the same mechanism as the appearance of energy gap in the theory of superconductivity. The idea can be put into a mathematical formulation utilizing a generalized Hartree-Fock approximation which regards real nucleons as quasi-particle excitations. We consider a simplified model of nonlinear four-fermion interaction which allows a  $\gamma_5$ -gauge group. An interesting consequence of the symmetry is that there arise automatically pseudoscalar zero-mass bound states of nucleon-antinucleon pair which may be regarded as an idealized pion. In addition, massive bound states of nucleon number zero and two are predicted in a simple approximation.

The theory contains two parameters which can be explicitly related to observed nucleon mass and the pion-nucleon coupling constant. Some paradoxical aspects of the theory in connection with the  $\gamma_5$  transformation are discussed in detail.

### I. INTRODUCTION

**I**N this paper we are going to develop a dynamical theory of elementary particles in which nucleons and mesons are derived in a unified way from a fundamental spinor field.<sup>1</sup> In basic physical ideas, it has thus the characteristic features of a compound-particle model, but unlike most of the existing theories, dynamical treatment of the interaction makes up an essential part of the theory. Strange particles are not yet considered.

The scheme is motivated by the observation of an interesting analogy between the properties of Dirac particles and the quasi-particle excitations that appear in the theory of superconductivity, which was originated with great success by Bardeen, Cooper, and Schrieffer,<sup>2</sup> and subsequently given an elegant mathematical formulation by Bogoliubov.<sup>3</sup> The characteristic feature of the BCS theory is that it produces an energy gap between the ground state and the excited states of a superconductor, a fact which has been confirmed experimentally. The gap is caused due to the fact that the attractive phonon-mediated interaction between electrons produces correlated pairs of electrons with opposite momenta and spin near the Fermi surface, and it takes a finite amount of energy to break this correlation.

Elementary excitations in a superconductor can be conveniently described by means of a coherent mixture of electrons and holes, which obeys the following

equations<sup>3,4</sup>:

$$\begin{aligned} E\psi_{p+} &= \epsilon_p \psi_{p+} + \phi \psi_{-p-}^*, \\ E\psi_{-p-}^* &= -\epsilon_p \psi_{-p-}^* + \phi \psi_{p+}, \end{aligned} \quad (1.1)$$

near the Fermi surface.  $\psi_{p+}$  is the component of the excitation corresponding to an electron state of momentum  $p$  and spin + (up), and  $\psi_{-p-}^*$  corresponding to a hole state of momentum  $p$  and spin +, which means an absence of an electron of momentum  $-p$  and spin - (down).  $\epsilon_p$  is the kinetic energy measured from the Fermi surface;  $\phi$  is a constant. There will also be an equation complex conjugate to Eq. (1), describing another type of excitation.

Equation (1) gives the eigenvalues

$$E_p = \pm (\epsilon_p^2 + \phi^2)^{\frac{1}{2}}. \quad (1.2)$$

The two states of this quasi-particle are separated in energy by  $2|E_p|$ . In the ground state of the system all the quasi-particles should be in the lower (negative) energy states of Eq. (2), and it would take a finite energy  $2|E_p| \geq 2|\phi|$  to excite a particle to the upper state. The situation bears a remarkable resemblance to the case of a Dirac particle. The four-component Dirac equation can be split into two sets to read

$$\begin{aligned} E\psi_1 &= \sigma \cdot p \psi_1 + m \psi_2, \\ E\psi_2 &= -\sigma \cdot p \psi_2 + m \psi_1, \\ E_p &= \pm (p^2 + m^2)^{\frac{1}{2}}, \end{aligned} \quad (1.3)$$

where  $\psi_1$  and  $\psi_2$  are the two eigenstates of the chirality operator  $\gamma_5 = \gamma_1 \gamma_2 \gamma_3 \gamma_4$ .

According to Dirac's original interpretation, the ground state (vacuum) of the world has all the electrons in the negative energy states, and to create excited states (with zero particle number) we have to supply an energy  $\geq 2m$ .

In the BCS-Bogoliubov theory, the gap parameter  $\phi$ , which is absent for free electrons, is determined essentially as a self-consistent (Hartree-Fock) representation of the electron-electron interaction effect.

\* Supported by the U. S. Atomic Energy Commission.

† Fulbright Fellow, on leave of absence from Instituto di Fisica dell' Università, Roma, Italy and Istituto Nazionale di Fisica Nucleare, Sezione di Roma, Italy.

<sup>1</sup> A preliminary version of the work was presented at the Midwestern Conference on Theoretical Physics, April, 1960 (unpublished). See also Y. Nambu, Phys. Rev. Letters 4, 380 (1960); and Proceedings of the Tenth Annual Rochester Conference on High-Energy Nuclear Physics, 1960 (to be published).

<sup>2</sup> J. Bardeen, L. N. Cooper, and J. R. Schrieffer, Phys. Rev. 106, 162 (1957).

<sup>3</sup> N. N. Bogoliubov, J. Exptl. Theoret. Phys. (U.S.S.R.) 34, 58, 73 (1958) [translation: Soviet Phys.-JETP 34, 41, 51 (1958)]; N. N. Bogoliubov, V. V. Tolmachev, and D. V. Shirkov, *A New Method in the Theory of Superconductivity* (Academy of Sciences of U.S.S.R., Moscow, 1958).

<sup>4</sup> J. G. Valatin, Nuovo cimento 7, 843 (1958).

One finds that

$$\phi \approx \omega \exp[-1/\rho], \quad (1.4)$$

where  $\omega$  is the energy bandwidth ( $\approx$  the Debye frequency) around the Fermi surface within which the interaction is important;  $\rho$  is the average interaction energy of an electron interacting with unit energy shell of electrons on the Fermi surface. It is significant that  $\phi$  depends on the strength of the interaction (coupling constant) in a nonanalytic way.

We would like to pursue this analogy mathematically. As the energy gap  $\phi$  in a superconductor is created by the interaction, let us assume that the mass of a Dirac particle is also due to some interaction between massless bare fermions. A quasi-particle in a superconductor is a mixture of bare electrons with opposite electric charges (a particle and a hole) but with the same spin; correspondingly a massive Dirac particle is a mixture of bare fermions with opposite chiralities, but with the same charge or fermion number. Without the gap  $\phi$  or the mass  $m$ , the respective particle would become an eigenstate of electric charge or chirality.

Once we make this analogy, we immediately notice further consequences of special interest. It has been pointed out by several people<sup>3,5-8</sup> that in a refined theory of superconductivity there emerge, in addition to the individual quasi-particle excitations, collective excitations of quasi-particle pairs. (These can alternatively be interpreted as moving states of bare electron pairs which are originally precipitated into the ground state of the system.) In the absence of Coulomb interaction, these excitations are phonon-like, filling the gap of the quasi-particle spectrum.

In general, they are excited when a quasi-particle is accelerated in the medium, and play the role of a backflow around the particle, compensating the change of charge localized on the quasi-particle wave packet. Thus these excitations are necessary consequences of the fact that individual quasi-particles are not eigenstates of electric charge, and hence their equations are not gauge invariant; whereas a complete description of the system must be gauge invariant. The logical connection between gauge invariance and the existence of collective states has been particularly emphasized by one of the authors.<sup>8</sup>

This observation leads to the conclusion that if a Dirac particle is actually a quasi-particle, which is only an approximate description of an entire system where chirality is conserved, then there must also exist collective excitations of bound quasi-particle pairs. The chirality conservation implies the invariance of the theory under the so-called  $\gamma_5$  gauge group, and from its nature one can show that the collective state must be a pseudoscalar quantity.

<sup>5</sup> D. Pines and J. R. Schrieffer, Nuovo cimento **10**, 496 (1958).

<sup>6</sup> P. W. Anderson, Phys. Rev. **110**, 827, 1900 (1958); **114**, 1002 (1959).

<sup>7</sup> G. Rickayzen, Phys. Rev. **115**, 795 (1959).

<sup>8</sup> Y. Nambu, Phys. Rev. **117**, 648 (1960).

It is perhaps not a coincidence that there exists such an entity in the form of the pion. For this reason, we would like to regard our theory as dealing with nucleons and mesons. The implication would be that the nucleon mass is a manifestation of some unknown primary interaction between originally massless fermions, the same interaction also being responsible for the binding of nucleon pairs into pions.

An additional support of the idea can be found in the weak decay processes of nucleons and pions which indicate that the  $\gamma_5$  invariance is at least approximately conserved, as will be discussed in Part II. There are some difficulties, however, that naturally arise on further examination.

Comparison between a relativistic theory and a non-relativistic, intuitive picture is often dangerous, because the former is severely restricted by the requirement of relativistic invariance. In our case, the energy-gap equation (4) depends on the energy density on the Fermi surface; for zero Fermi radius, the gap vanishes. The Fermi sphere, however, is not a relativistically invariant object, so that in the theory of nucleons it is not clear whether a formula like Eq. (4) could be obtained for the mass. This is not surprising, since there is a well known counterpart in classical electron theory that a finite electron radius is incompatible with relativistic invariance.

We avoid this difficulty by simply introducing a relativistic cutoff which takes the place of the Fermi sphere. Our framework does not yet resolve the divergence difficulty of self-energy, and the origin of such an effective cutoff has to be left as an open question.

The second difficulty concerns the mass of the pion. If pion is to be identified with the phonon-like excitations associated with a gauge group, its mass must necessarily be zero. It is true that in real superconductors the collective charge fluctuation is screened by Coulomb interaction to turn into the plasma mode, which has a finite "rest mass." A similar mechanism may be operating in the meson case too. It is possible, however, that the finite meson mass means that chirality conservation is only approximate in a real theory. From the evidence in weak interactions, we are inclined toward the second view.

The observation made so far does not yet give us a clue as to the exact mechanism of the primary interaction. Neither do we have a fundamental understanding of the isospin and strangeness quantum numbers, although it is easy to incorporate at least the isospin degree of freedom into the theory from the beginning. The best we can do here is to examine the various existing models for their logical simplicity and experimental support, if any. We will do this in Sec. 2, and settle for the moment on a nonlinear four-fermion interaction of the Heisenberg type. For reasons of simplicity in presentation, we adopt a model without isospin and strangeness degrees of freedom, and possessing complete  $\gamma_5$  invariance. Once the choice is made,

we can explore the whole idea mathematically, using essentially the formulation developed in reference 8. It is gratifying that the various field-theoretical techniques can be fully utilized. Section 3 will be devoted to introduction of the Hartree-Fock equation for nucleon self-energy, which will make the starting point of the theory. Then we go on to discuss in Sec. 4 the collective modes. In addition to the expected pseudoscalar "pion" states, we find other massive mesons of scalar and vector variety, as well as a scalar "deuteron." The coupling constants of these mesons can be easily determined. The relation of the pion to the  $\gamma_5$  gauge group will be discussed in Secs. 5 and 6.

The theory promises many practical consequences. For this purpose, however, it is necessary to make our model more realistic by incorporating the isospin, and allowing for a violation of  $\gamma_5$  invariance. But in doing so, there arise at the same time new problems concerning the mass splitting and instability. This refined model will be elaborated in Part II of this work, where we shall also find predictions about strong and weak interactions. Thus the general structure of the weak interaction currents modified by strong interactions can be treated to some degree, enabling one to derive the decay processes of various particles under simple assumptions. The calculation of the pion decay rate gives perhaps one of the most interesting supports of the theory. Results about strong interactions themselves are equally interesting. We shall find specific predictions about heavier mesons, which are in line with the recent theoretical expectations.

## II. THE PRIMARY INTERACTION

We briefly discuss the possible nature of the primary interaction between fermions. Lacking any radically new concepts, the interaction could be either mediated by some fundamental Bose field or due to an inherent nonlinearity in the fermion field. According to our postulate, these interactions must allow chirality conservation in addition to the conservation of nucleon number. The chirality  $X$  here is defined as the eigenvalue of  $\gamma_5$ , or in terms of quantized fields,

$$X = \int \bar{\psi} \gamma_4 \gamma_5 \psi d^3x. \quad (2.1)$$

The nucleon number is, on the other hand

$$N = \int \bar{\psi} \gamma_4 \psi d^3x. \quad (2.2)$$

These are, respectively, generators of the  $\gamma_5$ - and ordinary-gauge groups

$$\psi \rightarrow \exp[i\alpha \gamma_5] \psi, \quad \bar{\psi} \rightarrow \bar{\psi} \exp[i\alpha \gamma_5], \quad (2.3)$$

$$\psi \rightarrow \exp[i\alpha] \psi, \quad \bar{\psi} \rightarrow \bar{\psi} \exp[-i\alpha], \quad (2.4)$$

where  $\alpha$  is an arbitrary constant phase.

Furthermore, the dynamics of our theory would require that the interaction be attractive between particle and antiparticle in order to make bound-state formation possible. Under the transformation (2.3), various tensors transform as follows:

$$\begin{aligned} \text{Vector: } & i\bar{\psi} \gamma_\mu \psi \rightarrow i\bar{\psi} \gamma_\mu \psi, \\ \text{Axial vector: } & i\bar{\psi} \gamma_\mu \gamma_5 \psi \rightarrow i\bar{\psi} \gamma_\mu \gamma_5 \psi, \\ \text{Scalar: } & \bar{\psi} \psi \rightarrow \bar{\psi} \psi \cos 2\alpha + i\bar{\psi} \gamma_5 \psi \sin 2\alpha, \\ \text{Pseudoscalar: } & i\bar{\psi} \gamma_5 \psi \rightarrow i\bar{\psi} \gamma_5 \psi \cos 2\alpha - \bar{\psi} \psi \sin 2\alpha, \\ \text{Tensor: } & \bar{\psi} \sigma_{\mu\nu} \psi \rightarrow \bar{\psi} \sigma_{\mu\nu} \psi \cos 2\alpha + i\bar{\psi} \gamma_5 \sigma_{\mu\nu} \psi \sin 2\alpha. \end{aligned} \quad (2.5)$$

It is obvious that a vector or pseudovector Bose field coupled to the fermion field satisfies the invariance. The vector case would also satisfy the dynamical requirement since, as in the electromagnetic interaction, the forces would be attractive between opposite nucleon charges. The pseudovector field, on the other hand, does not meet the requirement as can be seen by studying the self-consistent mass equation discussed later.

The vector field looks particularly attractive since it can be associated with the nucleon number gauge group. This idea has been explored by Lee and Yang,<sup>9</sup> and recently by Sakurai.<sup>10</sup> But since we are dealing with strong interactions, such a field would have to have a finite observed mass in a realistic theory. Whether this is compatible with the invariance requirement is not yet clear. (Besides, if the bare mass of both spinor and vector field were zero, the theory would not contain any parameter with the dimensions of mass.)

The nonlinear fermion interaction seems to offer another possibility. Heisenberg and his co-workers<sup>11</sup> have been developing a comprehensive theory of elementary particles along this line. It is not easy, however, to gain a clear physical insight into their results obtained by means of highly complicated mathematical machinery.

We would like to choose the nonlinear interaction in this paper. Although this looks similar to Heisenberg's theory, the dynamical treatment will be quite different and more amenable to qualitative understanding.

The following Lagrangian density will be assumed ( $\hbar = c = 1$ ):

$$L = -\bar{\psi} \gamma_\mu \partial_\mu \psi + g_0 [(\bar{\psi} \psi)^2 - (\bar{\psi} \gamma_5 \psi)^2]. \quad (2.6)$$

The coupling parameter  $g_0$  is positive, and has dimensions [mass]<sup>-2</sup>. The  $\gamma_5$  invariance property of the interaction is evident from Eq. (2.5). According to the Fierz theorem, it is also equivalent to

$$-\frac{1}{2} g_0 [(\bar{\psi} \gamma_\mu \psi)^2 - (\bar{\psi} \gamma_\mu \gamma_5 \psi)^2]. \quad (2.7)$$

This particular choice of  $\gamma_5$ -invariant form was taken without a compelling reason, but has the advantage

<sup>9</sup> T. D. Lee and C. N. Yang, Phys. Rev. 98, 1501 (1955).

<sup>10</sup> J. J. Sakurai, Ann. Phys. 11, 1 (1960).

<sup>11</sup> W. Heisenberg, Z. Naturforsch. 14, 441 (1959). Earlier papers are quoted there.

that it can be naturally extended to incorporate isotopic spin.<sup>12</sup>

Unlike Heisenberg's case, we do not have any theory about the handling of the highly divergent singularities inherent in nonlinear interactions. So we will introduce, as an additional and independent assumption, an *ad hoc* relativistic cutoff or form factor in actual calculations. Thus the theory may also be regarded as an approximate treatment of the intermediate-boson model with a large effective mass.

As will be seen in subsequent sections, the nonlinear model makes mathematics particularly easy, at least in the lowest approximation, enabling one to derive many interesting quantitative results.

### III. THE SELF-CONSISTENT EQUATION FOR NUCLEON MASS

We will assume that all quantities we calculate here are somehow convergent, without asking the reason behind it. This will be done actually by introducing a suitable phenomenological cutoff.

Without specifying the interaction, let  $\Sigma$  be the unrenormalized proper self-energy part of the fermion, expressed in terms of observed mass  $m$ , coupling constant  $g$ , and cutoff  $\Lambda$ . A real Dirac particle will satisfy the equation

$$i\gamma \cdot p + m_0 + \Sigma(p, m, g, \Lambda) = 0 \quad (3.1)$$

for  $i\gamma \cdot p + m = 0$ . Namely

$$m - m_0 = \Sigma(p, m, g, \Lambda) \Big|_{i\gamma \cdot p + m = 0}. \quad (3.2)$$

The  $g$  will also be related to the bare coupling  $g_0$  by an equation of the type

$$g/g_0 = \Gamma(m, g, \Lambda). \quad (3.3)$$

Equations (3.1) and (3.2) may be solved by successive approximation starting from  $m_0$  and  $g_0$ . It is possible, however, that there are also solutions which cannot thus be obtained. In fact, there can be a solution  $m \neq 0$  even in the case where  $m_0 = 0$ , and moreover the symmetry seems to forbid a finite  $m$ .

This kind of situation can be most easily examined by means of the generalized Hartree-Fock procedure<sup>8,13</sup> which was developed before in connection with the theory of superconductivity. The basic idea is not new in field theory, and in fact in its simplest form the method is identical with the renormalization procedure of Dyson, considered only in a somewhat different context.

Suppose a Lagrangian is composed of the free and interaction part:  $L = L_0 + L_i$ . Instead of diagonalizing  $L_0$  and treating  $L_i$  as perturbation, we introduce the self-

energy Lagrangian  $L_s$ , and split  $L$  thus

$$\begin{aligned} L &= (L_0 + L_s) + (L_i - L_s) \\ &= L'_0 + L'_i. \end{aligned}$$

For  $L_s$  we assume quite general form (quadratic or bilinear in the fields) such that  $L'_0$  leads to linear field equations. This will enable one to define a vacuum and a complete set of "quasi-particle" states, each particle being an eigenmode of  $L'_0$ . Now we treat  $L'_i$  as perturbation, and determine  $L_s$  from the requirement that  $L'_i$  shall not yield additional self-energy effects. This procedure then leads to Eq. (3.2). The self-consistent nature of such a procedure is evident since the self-energy is calculated by perturbation theory with fields which are already subject to the self-energy effect.

In order to apply the method to our problem, let us assume that  $L_s = -m\bar{\psi}\psi$ , and introduce the propagator  $S_F^{(m)}(x)$  for the corresponding Dirac particle with mass  $m$ . In the lowest order, and using the two alternative forms Eqs. (2.6) and (2.7), we get for Eq. (3.2)

$$\Sigma = 2g_0 \left[ \text{Tr}S_F^{(m)}(0) - \gamma_5 \text{Tr}S_F^{(m)}(0)\gamma_5 - \frac{1}{2}\gamma_\mu \text{Tr}\gamma_\mu S_F^{(m)}(0) + \frac{1}{2}\gamma_\mu\gamma_5 \text{Tr}\gamma_\mu\gamma_5 S_F^{(m)}(0) \right] \quad (3.4)$$

in coordinate space.

This is quadratically divergent, but with a cutoff can be made finite. In momentum space we have

$$\Sigma = -\frac{8g_0 i}{(2\pi)^4} \int \frac{m}{p^2 + m^2 - i\epsilon} d^4 p F(p, \Lambda), \quad (3.5)$$

where  $F(p, \Lambda)$  is a cutoff factor. In this case the self-energy operator is a constant. Substituting  $\Sigma$  from Eq. (3.5), Eq. (3.2) gives ( $m_0 = 0$ )

$$m = -\frac{g_0 m i}{2\pi^4} \int \frac{d^4 p}{p^2 + m^2 - i\epsilon} F(p, \Lambda). \quad (3.6)$$

This has two solutions: either  $m = 0$ , or

$$1 = -\frac{g_0 i}{2\pi^4} \int \frac{d^4 p}{p^2 + m^2 - i\epsilon} F(p, \Lambda). \quad (3.7)$$

The first trivial one corresponds to the ordinary perturbative result. The second, nontrivial solution will determine  $m$  in terms of  $g_0$  and  $\Lambda$ .

If we evaluate Eq. (3.7) with a straight noninvariant cutoff at  $|p| = \Lambda$ , we get

$$\frac{\pi^2}{g_0 \Lambda^2} = \left( \frac{m^2}{\Lambda^2} + 1 \right)^{\frac{1}{2}} - \frac{m^2}{\Lambda^2} \ln \left[ \left( \frac{\Lambda^2}{m^2} + 1 \right)^{\frac{1}{2}} + \frac{\Lambda}{m} \right]. \quad (3.8)$$

If we use Eq. (3.5) with an invariant cutoff at  $p^2 = \Lambda^2$  after the change of path:  $p_0 \rightarrow ip_0$ , we get

$$\frac{2\pi^2}{g_0 \Lambda^2} = 1 - \frac{m^2}{\Lambda^2} \ln \left( \frac{\Lambda^2}{m^2} + 1 \right). \quad (3.9)$$

<sup>12</sup> This will be done in Part II.

<sup>13</sup> N. N. Bogoliubov, *Uspekhi Fiz. Nauk* **67**, 549 (1959) [translation: Soviet Phys.-Uspekhi **67**, 236 (1959)].

Since the right-hand side of Eq. (3.8) or (3.9) is positive and  $\leq 1$  for real  $\Lambda/m$ , the nontrivial solution exists only if

$$0 < 2\pi^2/g_0\Lambda^2 < 1. \quad (3.10)$$

Equation (3.9) is plotted in Fig. 1 as a function of  $m^2/\Lambda^2$ . As  $g_0\Lambda^2$  increases over the critical value  $2\pi^2$ ,  $m$  starts rising from 0. The nonanalytic nature of the solution is evident as  $m$  cannot be expanded in powers of  $g_0$ .

In the following we will assume that Eq. (3.10) is satisfied, so that the nontrivial solution exists. As we shall see later, physically this means that the nucleon-antinucleon interaction must be attractive ( $g_0 > 0$ ) and strong enough to cause a bound pair of zero total mass. In the BCS theory, the nontrivial solution corresponds to a superconductive state, whereas the trivial one corresponds to a normal state, which is not the true ground state of the superconductor. We may expect a similar situation to hold in the present case.

In this connection, it must be kept in mind that our solutions are only approximate ones. We are operating under the assumption that the corrections to them are not catastrophic, and can be appropriately calculated when necessary. If this does not turn out to be so for some solution, such a solution must be discarded. Later we shall indeed find this possibility for the trivial solution, but for the moment we will ignore such considerations.

Let us define then the vacuum corresponding to the two solutions. Let  $\psi^{(0)}$  and  $\psi^{(m)}$  be quantized fields satisfying the equations

$$\gamma_\mu \partial_\mu \psi^{(0)}(x) = 0, \quad (3.11a)$$

$$(\gamma_\mu \partial_\mu + m) \psi^{(m)}(x) = 0, \quad (3.11b)$$

$$\psi^{(0)}(x) = \psi^{(m)}(x) \quad \text{for } x_0 = 0. \quad (3.11c)$$

According to the standard procedure, we decompose the  $\psi$ 's into Fourier components:

$$\begin{aligned} \psi_\alpha^{(i)}(x) &= \frac{1}{V^{\frac{1}{2}}} \sum_{\substack{\mathbf{p}, s \\ p_0 = (\mathbf{p}^2 + m^2)^{\frac{1}{2}}}} [u_\alpha^{(i)}(\mathbf{p}, s) a^{(i)}(\mathbf{p}, s) e^{i\mathbf{p} \cdot x} \\ &\quad + v_\alpha^{*(i)}(\mathbf{p}, s) b^{(i)\dagger}(\mathbf{p}, s) e^{-i\mathbf{p} \cdot x}], \\ \psi_\alpha^{\dagger(i)}(x) &= \frac{1}{V^{\frac{1}{2}}} \sum_{\substack{\mathbf{p}, s \\ p_0 = (\mathbf{p}^2 + m^2)^{\frac{1}{2}}}} [u_\alpha^{(i)*}(\mathbf{p}, s) a^{(i)\dagger}(\mathbf{p}, s) \\ &\quad \times e^{-i\mathbf{p} \cdot x} + v_\alpha^{(i)}(\mathbf{p}, s) b^{(i)}(\mathbf{p}, s) e^{i\mathbf{p} \cdot x}], \end{aligned} \quad (3.12)$$

$i=0$  or  $m$ ,

where  $u_\alpha^{(i)}(\mathbf{p}, s)$ ,  $v_\alpha^{(i)}(\mathbf{p}, s)$  are the normalized spinor eigenfunctions for particles and antiparticles, with momentum  $\mathbf{p}$  and helicity  $s = \pm 1$ , and

$$\begin{aligned} \{a^{(i)}(\mathbf{p}, s), a^{(i)\dagger}(\mathbf{p}', s')\} \\ = \{b^{(i)}(\mathbf{p}, s), b^{(i)\dagger}(\mathbf{p}', s')\} = \delta_{pp'} \delta_{ss'}, \text{ etc.} \end{aligned} \quad (3.13)$$

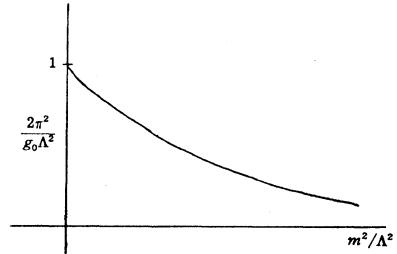


FIG. 1. Plot of the self-consistent mass equation (3.9).

The operator sets  $(a^{(0)}, b^{(0)})$  and  $(a^{(m)}, b^{(m)})$  are related by a canonical transformation because of Eq. (3.11c):

$$\begin{aligned} a^{(m)}(\mathbf{p}, s) &= \sum_{\alpha, s'} [u_\alpha^{(m)*}(\mathbf{p}, s) u_\alpha^{(0)}(\mathbf{p}, s') a^{(0)}(\mathbf{p}, s')] \\ &\quad + u_\alpha^{(m)*}(\mathbf{p}, s) v_\alpha^{(0)*}(-\mathbf{p}, s') b^{(0)\dagger}(-\mathbf{p}, s')], \\ b^{(m)}(\mathbf{p}, s) &= \sum_{\alpha, s'} [v_\alpha^{(m)*}(\mathbf{p}, s) v_\alpha^{(0)}(\mathbf{p}, s') b^{(0)}(\mathbf{p}, s')] \\ &\quad + v_\alpha^{(m)*}(\mathbf{p}, s) u_\alpha^{(0)*}(-\mathbf{p}, s') a^{(0)\dagger}(-\mathbf{p}, s'). \end{aligned} \quad (3.14)$$

Using Eq. (1.3), this is evaluated to give

$$\begin{aligned} a^{(m)}(\mathbf{p}, s) &= [\frac{1}{2}(1+\beta_p)]^{\frac{1}{2}} a^{(0)}(\mathbf{p}, s) \\ &\quad + [\frac{1}{2}(1-\beta_p)]^{\frac{1}{2}} b^{(0)\dagger}(-\mathbf{p}, s), \\ b^{(m)}(\mathbf{p}, s) &= [\frac{1}{2}(1+\beta_p)]^{\frac{1}{2}} b^{(0)}(\mathbf{p}, s) \\ &\quad - [\frac{1}{2}(1-\beta_p)]^{\frac{1}{2}} a^{(0)\dagger}(-\mathbf{p}, s), \\ \beta_p &= |\mathbf{p}| / (\mathbf{p}^2 + m^2)^{\frac{1}{2}}. \end{aligned} \quad (3.15)$$

The vacuum  $\Omega^{(0)}$  or  $\Omega^{(m)}$  with respect to the field  $\psi^{(0)}$  or  $\psi^{(m)}$  is now defined as

$$a^{(0)}(\mathbf{p}, s) \Omega^{(0)} = b^{(0)}(\mathbf{p}, s) \Omega^{(0)} = 0, \quad (3.16)$$

$$a^{(m)}(\mathbf{p}, s) \Omega^{(m)} = b^{(m)}(\mathbf{p}, s) \Omega^{(m)} = 0. \quad (3.16')$$

Both  $\psi^{(0)}$ ,  $\psi^{(0)}$  and  $\psi^{(m)}$ ,  $\psi^{(m)}$  applied to  $\Omega^{(0)}$  always create particles of mass zero, whereas the same applied to  $\Omega^{(m)}$  create particles of mass  $m$ .

From Eqs. (3.15) and (3.16) we obtain

$$\begin{aligned} \Omega^{(m)} &= \prod_{\mathbf{p}, s} \{[\frac{1}{2}(1+\beta_p)]^{\frac{1}{2}} \\ &\quad - [\frac{1}{2}(1-\beta_p)]^{\frac{1}{2}} a^{(0)\dagger}(\mathbf{p}, s) b^{(0)\dagger}(-\mathbf{p}, s)\} \Omega^{(0)}. \end{aligned} \quad (3.17)$$

Thus  $\Omega^{(m)}$  is, in terms of zero-mass particles, a superposition of pair states. Each pair has zero momentum, spin and nucleon number, and carries  $\pm 2$  units of chirality, since chirality equals minus the helicity  $s$  for massless particles.

Let us calculate the scalar product  $(\Omega^{(0)}, \Omega^{(m)})$  from Eq. (3.15):

$$\begin{aligned} (\Omega^{(0)}, \Omega^{(m)}) &= \prod_{\mathbf{p}, s} [\frac{1}{2}(1+\beta_p)]^{\frac{1}{2}} \\ &= \exp \left\{ \sum_{\mathbf{p}, s} \frac{1}{2} \ln [\frac{1}{2}(1+\beta_p)] \right\}. \end{aligned} \quad (3.18)$$

For large  $p$ ,  $\beta_p \sim 1 - m^2/2p^2$ , so that the exponent

diverges as  $V\pi m^2 \int d\mathbf{p}/(2\pi)^3$  ( $V$ =normalization volume). Hence

$$(\Omega^{(0)}, \Omega^{(m)}) = 0. \quad (3.19)$$

It is easy to see that any two states  $\Psi^{(0)}$  and  $\Psi^{(m)}$ , obtained by applying a finite number of creation operators on  $\Omega^{(0)}$  and  $\Omega^{(m)}$  respectively, are also orthogonal.

Thus the two "worlds" based on  $\Omega^{(0)}$  and  $\Omega^{(m)}$  are physically distinct and outside of each other. No interaction or measurement, in the usual sense, can bridge them in finite steps.

What is the energy difference of the two vacua? Since both are Lorentz invariant states, the difference can only be either zero or infinity. Using the expression

$$\begin{aligned} H^{(m)} &= \sum_{\mathbf{p}, s} (\mathbf{p}^2 + m^2)^{\frac{1}{2}} \{ a^{(m)\dagger}(\mathbf{p}, s) a^{(m)}(\mathbf{p}, s) \\ &\quad - b^{(m)}(\mathbf{p}, s) b^{(m)\dagger}(\mathbf{p}, s) \}, \\ H^{(0)} &= \sum_{\mathbf{p}, s} |\mathbf{p}| \{ a^{(0)\dagger}(\mathbf{p}, s) a^{(0)}(\mathbf{p}, s) \\ &\quad - b^{(0)}(\mathbf{p}, s) b^{(0)\dagger}(\mathbf{p}, s) \}, \end{aligned} \quad (3.20)$$

we get for the respective energies

$$E^{(m)} - E^{(0)} = -2 \sum_{\mathbf{p}} [(\mathbf{p}^2 + m^2)^{\frac{1}{2}} - |\mathbf{p}|], \quad (3.21)$$

which is negative and quadratically divergent. So  $\Omega^{(m)}$  may be called the "true" ground state, as was expected.

There remains finally the question of  $\gamma_5$  invariance. The original Hamiltonian allowed two conservations  $X$  and  $N$ , Eqs. (2.1) and (2.2). Both  $\Omega^{(0)}$  and  $\Omega^{(m)}$  belong to  $N=0$ , and their elementary excitations carry  $N=\pm 1$ . In the case of  $X$ , the same is true for the space  $\Omega^{(0)}$ , but  $\Omega^{(m)}$  as well as its elementary excitations are not eigenstates of  $X$ , as is clear from the foregoing results. If the latter solution is to be a possibility, there must be an infinite degeneracy with respect to the quantum number  $X$ . A ground state will be in general a linear combination of degenerate states with different  $X=0, \pm 2, \dots$ :

$$\Omega^{(m)} = \sum_{n=-\infty}^{\infty} C_{2n} \Omega_{2n}^{(m)}. \quad (3.22)$$

Equation (3.17) is in fact a particular case of this. The  $\gamma_5$ -gauge transformation Eq. (2.3) induces the change

$$\begin{aligned} a^{(0)}(\mathbf{p}, \pm 1) &\rightarrow e^{\mp i\alpha} a^{(0)}(\mathbf{p}, \pm 1), \\ b^{(0)}(\mathbf{p}, \pm 1) &\rightarrow e^{\mp i\alpha} b^{(0)}(\mathbf{p}, \pm 1), \\ a^{(0)\dagger}(\mathbf{p}, \pm 1) &\rightarrow e^{\pm i\alpha} a^{(0)\dagger}(\mathbf{p}, \pm 1), \\ b^{(0)\dagger}(\mathbf{p}, \pm 1) &\rightarrow e^{\pm i\alpha} b^{(0)\dagger}(\mathbf{p}, \pm 1), \end{aligned} \quad (3.23)$$

and the coefficients of Eq. (3.22) become

$$C_{2n} \rightarrow e^{-2n i\alpha} C_{2n}. \quad (3.24)$$

In particular

$$\begin{aligned} \Omega^{(m)} &\rightarrow \Omega_{\alpha}^{(m)} \\ &= \exp[-i\alpha X] \Omega^{(m)} \\ &= \prod_{\mathbf{p}, \pm} \{ [\frac{1}{2}(1+\beta_p)]^{\frac{1}{2}} - [\frac{1}{2}(1-\beta_p)]^{\frac{1}{2}} \\ &\quad \times e^{\pm 2i\alpha} a^{(0)\dagger}(\mathbf{p}, \pm) b^{(0)\dagger}(-\mathbf{p}, \pm) \} \Omega^{(0)}. \end{aligned} \quad (3.25)$$

The Dirac equation (3.11b), at the same time, is transformed into

$$[\gamma_{\mu} \partial_{\mu} + m \cos 2\alpha + im\gamma_5 \sin 2\alpha] \psi = 0. \quad (3.26)$$

The moral of this is that the self-consistent self-energy  $\Sigma$  is determined only up to a  $\gamma_5$  transformation. This can be easily verified from Eq. (3.4), in which the second term on the right-hand side is nonvanishing when a propagator corresponding to Eq. (3.26) is used. Although Eq. (3.26) seems to violate parity conservation, it is only superficially so since  $\Omega_{\alpha}^{(m)}$  is now not an eigenstate of parity. We could alternatively say that the parity operator undergoes transformation together with the mass operator. Despite the odd form of the equation (3.26), there is no change in the physical predictions of the theory. We shall see more of this later.

Let us calculate, as before, the scalar product of  $\Omega_{\alpha}^{(m)}$  and  $\Omega_{\alpha'}^{(m)}$ . From Eqs. (3.17) and (3.25) we get

$$\begin{aligned} &(\Omega_{\alpha}^{(m)}, \Omega_{\alpha'}^{(m)}) \\ &= \prod_{\mathbf{p}, \pm} [\frac{1}{2}(1+\beta_p) - e^{\pm 2i(\alpha'-\alpha)} \frac{1}{2}(1-\beta_p)] \\ &= \prod_{\mathbf{p}, \pm} [1 + (e^{\pm 2i(\alpha'-\alpha)} - 1) \frac{1}{2}(1-\beta_p)] \\ &= \exp \left\{ \sum_{\mathbf{p}, \pm} \ln [1 + (e^{\pm 2i(\alpha'-\alpha)} - 1) \frac{1}{2}(1-\beta_p)] \right\}. \end{aligned} \quad (3.27)$$

For large  $|\mathbf{p}|$ , the exponent goes like

$$\frac{V}{(2\pi)^3} \sum_{\pm} (e^{\pm 2i(\alpha'-\alpha)} - 1) \int \frac{m^2}{4\mathbf{p}^2} d^3 p.$$

The integral is again divergent. Hence

$$\begin{aligned} (\Omega_{\alpha}^{(m)}, \Omega_{\alpha'}^{(m)}) &= (\Omega^{(m)}, \exp[-i(\alpha'-\alpha)X] \Omega^{(m)}) \\ &= 0, \quad \alpha' \neq \alpha \pmod{2\pi}, \end{aligned} \quad (3.28)$$

and, of course

$$(\Omega^{(0)}, \Omega_{\alpha}^{(m)}) = 0. \quad (3.28')$$

We can evaluate  $(\Omega_{\alpha}^{(m)}, \Omega_{\alpha'}^{(m)})$  alternatively from Eqs. (3.22) and (3.24). Then

$$\sum_{m=-\infty}^{\infty} |C_{2m}|^2 e^{2n i(\alpha-\alpha')} = 0, \quad \alpha \neq \alpha' \pmod{2\pi}, \quad (3.29)$$

implying that

$$|C_0| = |C_{\pm 2}| = |C_{\pm 4}| = \dots = C. \quad (3.30)$$

Thus there is an infinity of equivalent worlds described by  $\Omega_{\alpha}^{(m)}$ ,  $0 \leq \alpha < 2\pi$ . The states  $\Omega_{2n}$  of Eq. (3.22) are then expressed in terms of  $\Omega_{\alpha}^{(m)}$  as

$$C_{2n} \Omega_{2n}^{(m)} = \frac{1}{2\pi} \int_0^{2\pi} e^{2n i\alpha} \Omega_{\alpha}^{(m)} d\alpha, \quad (3.31)$$

which form another orthogonal set. Since the original total  $H$  commutes with  $X$ , it will have no matrix elements connecting different "worlds." Moreover, as

was the case with  $\Omega^{(m)}$  and  $\Omega^{(0)}$ , no finite measurement can induce similar transitions. This is a kind of superselection rule, which effectively avoids the apparent degeneracy to show up as physical effects.<sup>14</sup> The usual description of the world by means of  $\Omega^{(m)}$  and ordinary Dirac particles must be regarded as only the most convenient one.

We still are left with some paradoxes. The  $X$  conservation implies the existence of a conserved  $X$  current:

$$j_{\mu 5} = i\bar{\psi}\gamma_{\mu}\gamma_5\psi, \quad (3.32)$$

$$\partial_{\mu}j_{\mu 5} = 0, \quad (3.32')$$

which can readily be verified from Eq. (2.6). On the other hand, for a massive Dirac particle the continuity equation is not satisfied:

$$\partial_{\mu}\bar{\psi}^{(m)}\gamma_{\mu}\gamma_5\psi^{(m)} = 2m\bar{\psi}^{(m)}\gamma_5\psi^{(m)}. \quad (3.33)$$

If a massive Dirac particle has to be a real eigenstate of the system, how can this be reconciled? The answer would be that the  $X$ -current operator taken between real one-nucleon states should not be given simply by  $i\gamma_{\mu}\gamma_5$  because of the "radiative corrections." We expect instead

$$\langle p' | j_{\mu 5} | p \rangle = \bar{u}(p')X_{\mu}(p', p)u(p), \quad (3.34)$$

where the renormalized quantity  $X_{\mu 5}$  should be, from relativistic invariance grounds, of the form

$$X_{\mu}(p', p) = F_1(q^2)i\gamma_{\mu}\gamma_5 + F_2(q^2)\gamma_5q_{\mu}, \quad (3.35)$$

$$q = p' - p, \quad q^2 = p'^2 - p^2 = -m^2.$$

The continuity equation (3.32'), together with Eq. (3.33), further reduces this to

$$F_1 = F_2 q^2 / 2m \equiv F, \quad (3.36)$$

$$X_{\mu}(p', p) = F(q^2) \left( i\gamma_{\mu}\gamma_5 + \frac{2m\gamma_5q_{\mu}}{q^2} \right).$$

The real nucleon is not a point particle. Its  $X$ -current (3.36) is provided with the dramatic "anomalous" term.

To understand the physical meaning of the anomalous term, we have to make use of the dispersion relations. The form factors  $F_1$  and  $F_2$  will, in general, satisfy dispersion relations of the form

$$F_i(q^2) = F_i(0) - \frac{q^2}{\pi} \int \frac{\text{Im}F_i(-\kappa^2)}{(q^2 + \kappa^2 - i\epsilon)\kappa^2} d\kappa^2, \quad (3.37)$$

assuming one subtraction. Each singularity at  $\kappa^2$  corresponds to some physical intermediate state. Thus if  $F(0) \neq 0$ , Eq. (3.36) indicates that there is a pole at  $q^2 = 0$  for  $F_2$  (and no subtraction), which means in turn that there is an isolated intermediate state of zero mass.

<sup>14</sup>This was discussed by R. Haag, Kgl. Danske Videnskab. Selskab, Mat.-fys. Medd. **29**, No. 12 (1955). See also L. van Hove, Physica **18**, 145 (1952).

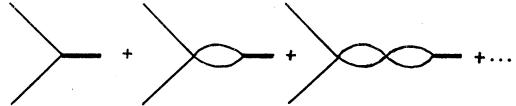


FIG. 2. Graphs corresponding to the Bethe-Salpeter equation in "ladder" approximation. The thick line is a bound state.

To see its nature, we take a time-like  $q$  in its own rest frame and go to the limit  $q^2 \rightarrow 0$ . The anomalous term has then only the time component, and is proportional to the amplitude for creation of a nucleon pair in a  $J=0^-$  state. Hence the zero mass state must have the same property as this pair. It belongs to nucleon number zero, so that we may call it a zero-mass pseudoscalar meson. In order for a  $\gamma_5$ -invariant Hamiltonian such as Eq. (2.6) to allow massive nucleon states and a nonvanishing  $X$  current for  $q=0$ , it is therefore necessary to have at the same time pseudoscalar zero-mass mesons coupled with the nucleons. Since we did not have such mesons in the theory, they must be regarded as secondary products, i.e., bound states of nucleon pairs. This conclusion would not hold if in Eq. (3.36)  $F(q^2) = O(q^2)$  near  $q^2 = 0$ . A nucleon then would have always  $X=0$ . Such a possibility cannot be excluded. We will show, however, that the pseudoscalar zero-mass bound states do follow explicitly, once we assume the nontrivial solution of the self-energy equation.

#### IV. THE COLLECTIVE STATES

From the general discussion of Secs. 2 and 3, we may expect the existence of collective states of the fundamental field which would manifest themselves as stable or unstable particles. In particular we have argued that, as a consequence of the  $\gamma_5$  invariance, a pseudoscalar zero-mass state must exist. We want now to discuss the problem in detail, trying to determine the mass spectrum of the collective excitations (at least its general features) and the strength of their coupling with the nucleons. These states must be considered as a direct effect of the same primary interaction which produces the mass of the nucleon, which itself is a collective effect. We will study the bound-state problem through the use of the Bethe-Salpeter equation, taking into account explicitly the self-consistency conditions. We first verify in the following the existence of the zero-mass pseudoscalar state.

The Bethe-Salpeter equation for a bound pair  $B$  deals with the amplitude

$$\Phi(x, y) = \langle 0 | T(\psi(x)\bar{\psi}(y)) | B \rangle. \quad (4.1)$$

As is well known, the equation is relatively easy to handle in the ladder approximation. In our case we have a four-spinor point interaction and the analog of the "ladder" approximation would be the iteration of the simplest closed loop (see Fig. 2) in which all lines represent dressed particles. We introduce the vertex function

$\Gamma$  related to  $\Phi$  by

$$\Phi(p) = S_F^{(m)}(p + \frac{1}{2}q)\Gamma(p + \frac{1}{2}q, p - \frac{1}{2}q)S_F^{(m)}(p - \frac{1}{2}q). \quad (4.2)$$

All we have to do then is to set up the integral equation generated by the chain of diagrams, looking for solutions having the symmetry properties of a pseudoscalar state. This means that our solutions must be proportional to  $\gamma_5$ . This requirement makes only the pseudoscalar and axial vector part of the interaction contribute to the integral equation. We have

$$\begin{aligned} & \Gamma(p + \frac{1}{2}q, p - \frac{1}{2}q) \\ &= \frac{2ig_0}{(2\pi)^4} \gamma_5 \int \text{Tr}[\gamma_5 S_F^{(m)}(p' + \frac{1}{2}q) \\ & \quad \times \Gamma(p' + \frac{1}{2}q, p' - \frac{1}{2}q) S_F^{(m)}(p' - \frac{1}{2}q)] d^4 p' \\ & \quad - \frac{ig_0}{(2\pi)^4} \gamma_5 \gamma_\mu \int \text{Tr}[\gamma_5 \gamma_\mu S_F^{(m)}(p' + \frac{1}{2}q) \\ & \quad \times \Gamma(p' + \frac{1}{2}q, p' - \frac{1}{2}q) S_F^{(m)}(p' - \frac{1}{2}q)] d^4 p'. \end{aligned} \quad (4.3)$$

For the moment let us ignore the pseudovector term on the right-hand side. It then follows that the equation has a constant solution  $\Gamma = C\gamma_5$  if  $q^2 = 0$ . To see this, first observe that for the special case  $q = 0$ , Eq. (4.3) reduces to

$$1 = -\frac{8ig_0}{(2\pi)^4} \int \frac{d^4 p}{p^2 + m^2 - i\epsilon}, \quad (4.4)$$

which is nothing but the self-consistency condition (3.7), provided that the same cutoff is applied. Since the pseudoscalar term of Eq. (4.3) gives a function of  $q^2$  only, the same condition remains true as long as  $q^2 = 0$ .

When the pseudovector term is included, we have still the same eigenvalue  $q^2 = 0$  with a solution of the form  $\Gamma = C\gamma_5 + iD\gamma_5 \gamma \cdot q$ , which is not difficult to verify (see Appendix).

We now add some remarks. First, the bound state amplitude for this solution spreads in space over a region of the order of the fermion Compton wavelength  $1/m$  because of Eq. (4.2), making the zero-mass particle only partially localizable. We want also to stress the role played by the  $\gamma_5$  invariance in the argument. We had in fact already inferred the existence of the pseudoscalar particle from relativistic and  $\gamma_5$  invariance alone, and at first sight the same result seems to follow now essentially from the self-consistency equation. However, we must notice that only the scalar term of the Lagrangian appears in this equation while only the pseudoscalar part contributes in the Bethe-Salpeter equation. It is because of the  $\gamma_5$ -invariant Lagrangian that the Bethe-Salpeter equation can be reduced to the self-consistency condition.

Along the same line we could try to see whether other bound states exist in the "ladder" approximation. However, besides calculating the spectrum, it is also im-

portant to determine the interaction properties of these collective states with the fermions. For this purpose the study of the two-“nucleon” scattering amplitude appears much more suitable, as we shall realize after the following remark. Once we have recognized that in the ladder approximation the collective states would appear as real stable particles, we must expect to the same degree of approximation poles in the scattering matrix of two nucleons corresponding to the possibility of the virtual exchange of these particles. For definiteness we shall refer again as an example to the pseudoscalar zero-mass particle. Let us indicate by  $J_p(q)$  the analytical expression corresponding to the graph whose iteration produces the bound state [Fig. 3(a)]. We construct next the scattering matrix generated by the exchange of all possible simple chains built with this element. This means that we consider the set of diagrams in Fig. 3(b). The series is easily evaluated and we obtain

$$\frac{1}{2g_0 i\gamma_5} \frac{1}{1 - J_p(q)} i\gamma_5, \quad (4.5)$$

where the  $\gamma_5$ 's refer to the pairs (1,1') and (2,2'), respectively. The meaning of this result is clear: because of the self-consistent equation  $J_p(0) = 1$ , Eq. (4.5) is equivalent to a phenomenological exchange term where the intermediate particle is our pseudoscalar massless boson (Fig. 4). The coupling constant  $G$  can now be evaluated by straightforward comparison. Before doing this calculation we need the explicit expression of  $J_p(q)$ . Using the ordinary rules for diagrams, we have

$$\begin{aligned} J_p(q) &= -\frac{2ig_0}{(2\pi)^4} \\ & \times \int \frac{4(m^2 + p^2) - q^2}{[(p + \frac{1}{2}q)^2 + m^2][(p - \frac{1}{2}q)^2 + m^2]} d^4 p. \end{aligned} \quad (4.6)$$

It is however more convenient to rewrite  $J_p$  in the form of a dispersive integral, and if we forget for a moment that it is a divergent expression, a simple manipulation gives

$$J_p(q) = \frac{g_0}{4\pi^2} \int_{4m^2}^{\Lambda^2} \frac{\kappa^2 (1 - 4m^2/\kappa^2)^{\frac{1}{2}}}{q^2 + \kappa^2} d\kappa^2. \quad (4.6')$$

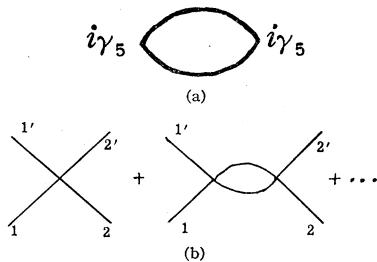


FIG. 3. The bubble graph for  $J_p$  and the scattering matrix generated by it.

In order for this expression to be meaningful, a new cutoff  $\Lambda$  must be introduced. There is no simple relation between this and the previous cutoffs. The dispersive form is more comfortable to handle and accordingly we shall reformulate the self-consistent condition  $J_P(0)=1$ , or

$$1 = \frac{g_0}{4\pi^2} \int_{4m^2}^{\Lambda^2} (1 - 4m^2/\kappa^2)^{\frac{1}{2}} d\kappa^2. \quad (4.7)$$

It may be of interest to remark at this point that Eq. (4.7) can be obtained also if we think of our theory as a theory with intermediate pseudoscalar boson in the limit of infinite boson mass. We are now in a position to evaluate the phenomenological coupling constant  $G$ . From Eqs. (4.6') and (4.7) we have

$$J_P(q^2) = 1 - q^2 \frac{g_0}{4\pi^2} \int_{4m^2}^{\Lambda^2} \frac{(1 - 4m^2/\kappa^2)^{\frac{1}{2}}}{q^2 + \kappa^2} d\kappa^2, \quad (4.8)$$

which leads immediately to the result

$$\frac{G_P^2}{4\pi} = 2\pi \left[ \int_{4m^2}^{\Lambda^2} \frac{(1 - 4m^2/\kappa^2)^{\frac{1}{2}}}{\kappa^2} d\kappa^2 \right]^{-1}. \quad (4.9)$$

This equation is interesting since it establishes a connection between the phenomenological constant  $G_P$  and the cutoff independently of the value of the fundamental coupling  $g_0$ . This fact exhibits the purely dynamical origin of the phenomenological coupling  $G_P$ . Actually  $g_0$  is buried in the value of the mass  $m$ .

So far we have exploited only the  $\gamma_5$  vertex. What happens then if the scalar part is iterated to form chains of bubbles similar to those we have already discussed? The procedure just explained can be followed again, and a quantity  $J_S(q)$  can be defined similarly with the result

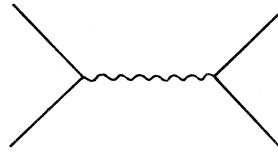
$$J_S(q) = \frac{g_0}{4\pi^2} \int_{4m^2}^{\Lambda^2} \frac{(\kappa^2 - 4m^2)(1 - 4m^2/\kappa^2)^{\frac{1}{2}}}{q^2 + \kappa^2} d\kappa^2. \quad (4.10)$$

It is immediately seen that because of Eq. (4.7)

$$J_S(-4m^2) = 1, \quad (4.11)$$

which causes a new pole to appear in the  $S$  matrix for  $q^2 = -4m^2$ . This means that we have another collective state of mass  $2m$ , parity + and spin 0! We observe that it is necessary to assume the same cutoff as in the pseudoscalar case in order that this result may be obtained. The choice of the same cutoff in both cases seems to be suggested by the  $\gamma_5$  invariance as will be seen later. We also notice the peculiar symmetry existing between the pseudoscalar and the scalar state: the first has zero mass and binding energy  $2m$ , while the opposite is true for the scalar particle. So in the bound-state picture the scalar particle would not be a true bound state and should be, rather, interpreted as a

FIG. 4. The equivalent phenomenological one-meson exchange graph.



correlated exchange of pairs in the scattering process.<sup>15</sup> The "nucleon-nucleon" forces induced by the exchange of the scalar particle are, of course, of rather short range. The general physical implications of these results will be discussed more thoroughly later.

The phenomenological coupling constant  $G_S$  for the scalar meson is given by

$$\frac{G_S^2}{4\pi} = 2\pi \left[ \int_{4m^2}^{\Lambda^2} \frac{(1 - 4m^2/\kappa^2)^{\frac{1}{2}}}{(\kappa^2 - 4m^2)} d\kappa^2 \right]^{-1}. \quad (4.12)$$

Let us next turn to the vector state generated by iteration of the vector interaction. In this case we obtain for each "bubble" a tensor

$$\begin{aligned} J_{V\mu\nu} &= (\delta_{\mu\nu} - q_\mu q_\nu/q^2) J_V, \\ J_V &= -\frac{g_0}{4\pi^2} \frac{q^2}{3} \int_{4m^2}^{\Lambda^2} \frac{d\kappa^2}{q^2 + \kappa^2} \\ &\times \left( 1 + \frac{2m^2}{\kappa^2} \right) (1 - 4m^2/\kappa^2)^{\frac{1}{2}}. \end{aligned} \quad (4.13)$$

Perhaps a remark is in order here regarding the evaluation of  $J_V$ . It suffers from an ambiguity of subtraction well known in connection with the photon self-energy problem. The above result is of the conventional gauge invariant form, which we take to be the proper choice.

Equation (4.13) leads to the scattering matrix

$$\frac{1}{g_0} \left[ \frac{1}{1 - J_V} \gamma_\mu - \gamma^\cdot q \frac{J_V}{(1 - J_V)q^2} \gamma^\cdot q \right], \quad (4.14)$$

where the second term is, of course, effectively zero. It can be easily seen that the denominator can produce a pole below  $4m^2$  for sufficiently small  $\Lambda^2$ . In fact, from Eqs. (4.7) and (4.13), we find

$$(8/3)m^2 < \mu_V^2. \quad (4.15)$$

The coupling constant is given by

$$\frac{G_V^2}{4\pi} = 3\pi \left[ \int_{4m^2}^{\Lambda^2} d\kappa^2 \frac{\kappa^2 + 2m^2}{(\kappa^2 - \mu_V^2)^2} (1 - 4m^2/\kappa^2)^{\frac{1}{2}} \right]^{-1}. \quad (4.16)$$

It must be noted that the mass of the vector meson now depends on the cutoff, unlike the previous two cases.

Finally we are left with the pseudovector state. We

<sup>15</sup> Of course this and other heavy mesons will in general become unstable in higher order approximation, which is beyond the scope of the present paper.

find for the bubble<sup>16</sup>

$$\begin{aligned} J_{A\mu\nu} &= -J_{V\mu\nu} + J_{A'}\delta_{\mu\nu}, \\ J_{A'} &= \frac{g_0 m^2}{2\pi^2} \int_{4m^2}^{\Lambda^2} \frac{dk^2}{q^2 + k^2} (1 - 4m^2/k^2)^{\frac{1}{2}}. \end{aligned} \quad (4.17)$$

In view of the self-consistency condition (4.7), it can be seen that this does not produce a pole of the scattering matrix for  $-q^2 < 4m^2$ , corresponding to a pseudovector meson.

So far we have considered only iterations of the same kind of interactions. In the ladder approximation there is actually a coupling between pseudoscalar and pseudovector interactions as was explicitly considered in Eq. (4.3). However, the coupling between scalar and vector interactions vanish because of the Furry's theorem.

This coupling of pseudoscalar and pseudovector interactions does not change the pion pole of the scattering matrix, but it affects the coupling of the pion to the nucleon since a chain of the pseudoscalar can join the external nucleon with an axial vector interaction. In other words, the pion-nucleon coupling is in general a mixture of pseudoscalar and derivative pseudovector types (Appendix).

We would like to inject here a remark concerning the trivial solution of the self-energy equation, against which we had no decisive argument. So let us also try to apply our scattering formula to this solution. For the pseudoscalar state we now find  $J_P(q=0) > 1$ , provided that the cutoff  $\Lambda$  is kept fixed and  $m$  is set equal to zero in Eq. (4.6'). (The pseudovector interference vanishes.) In other words, there will be a pole for some  $q^2 > 0$  ( $\mu^2 < 0$ ). This is again a supporting evidence that the trivial solution could be unstable, capable of decaying by emitting such mesons. The final answer, however, depends on the exact nature of the cutoff.

Finally we would like to discuss the nucleon-nucleon scattering in the same spirit and approximation as for the nucleon-antinucleon scattering. In order to make a correspondence with the previous cases, it is convenient to rewrite the Hamiltonian in the following way:

$$\begin{aligned} H_1 &= -g_0 [\bar{\psi}\psi\bar{\psi}\psi^c - \bar{\psi}\gamma_5\psi\bar{\psi}\gamma_5\psi^c] \\ &= \frac{1}{2}g_0 [\bar{\psi}\gamma_\mu\psi^c\bar{\psi}\gamma_\mu\psi - \bar{\psi}\gamma_\mu\gamma_5\psi^c\bar{\psi}\gamma_5\gamma_\mu\psi] \\ &= -\frac{1}{2}g_0 [\bar{\psi}\gamma_\mu C\bar{\psi}\psi^c C^{-1}\gamma_\mu\psi - \bar{\psi}\gamma_\mu\gamma_5 C\bar{\psi}\psi^c C^{-1}\gamma_\mu\gamma_5\psi], \end{aligned} \quad (4.18)$$

where  $\psi^c$ ,  $\bar{\psi}^c$  are the charge-conjugate fields.

The last form of Eq. (4.18) is suitable for our purpose. We note first that the vector part of the interaction is identically zero because of the anticommutativity of  $\psi$ . Thus only the pseudovector part survives. A "bubble" made of this interaction then is seen to give rise to the same integral  $J_A$ , Eq. (4.17). Since the interfering pseudoscalar interaction is missing in the present case,

<sup>16</sup> We meet here again the problem of subtraction. Our choice follows naturally from comparison with the vector case, and is consistent with Eq. (3.33).

TABLE I. Mass spectrum.

Nucleon number	Mass $\mu$	Spin-parity	Spectroscopic notation
0	0	$0^-$	$^1S_0$
0	$2m$	$0^+$	$^3P_0$
0	$(8/3)m^2 < \mu^2$	$1^-$	$^3P_1$
$\pm 2$	$2m^2 < \mu^2$	$0^+$	$^1S_0$

we get the complete scattering matrix by iterating  $J_A$ :

$$\begin{aligned} &- \gamma_\mu\gamma_5 C \left[ \frac{\delta_{\mu\nu} - q_\mu q_\nu/q^2}{1 - J_A} + \frac{q_\mu q_\nu/q^2}{1 - J_{A'}} \right] C^{-1} \gamma\gamma_5 \\ &= \gamma_\mu\gamma_5 C \frac{1}{1 - J_A} C^{-1} \gamma_\mu\gamma_5 \\ &\quad + \gamma \cdot q \gamma_5 C \frac{J_V/q^2}{(1 - J_{A'})(1 - J_A)} C^{-1} \gamma \cdot q \gamma_5, \end{aligned} \quad (4.19)$$

$$J_A \equiv J_{A'} - J_V.$$

The first term, corresponding to a scattering in the  $J=1^-$  state, does not have a pole. The second term can have one below  $4m^2$  for  $1 = J_{A'}$ . With Eqs. (4.7) and (4.17), this determines the mass  $\mu_D$ :

$$2m^2 < \mu_D^2. \quad (4.20)$$

In this second term of the scattering matrix, the wave function is proportional to  $C\gamma \cdot q \gamma_5$ , so that the bound state behaves like a scalar "deuteron" (a singlet  $S$  state). The residue of the pole determines the nucleon-"deuteron" coupling constant (derivative)  $G_D^2$ , which is positive as it should be.

Table I summarizes the main results of this section. Although our approximation is a very crude one, we believe that it reflects the real situation at least qualitatively, because all the results are understandable in simple physical terms. Thus in the nonrelativistic sense, our Hamiltonian contains spin-independent attractive scalar and vector interactions plus a spin-dependent axial vector interaction between a particle and an antiparticle. Between particles, the vector part turns into a repulsion. Table I is just what we expect for the level ordering from this consideration.

## V. PHENOMENOLOGICAL THEORY AND $\gamma_5$ INVARIANCE

In the previous section special subsets of diagrams were taken into account, and the existence of various boson states was established, together with their couplings with the nucleons. As was discussed there, we can reasonably expect that these results are essentially correct in spite of the very simple approximations. Because the bosons have in general small masses (compared to the unbound nucleon states), they will play important roles in the dynamics of strong interactions at least at energies comparable to these masses.

Thus if we are willing to accept the conclusions of our lowest order approximation, what we should do then is to study the dynamics of systems consisting of nucleons and the different kinds of bosons which all together represent the primary manifestation of the fundamental interaction. These particles will be now assumed to interact via their phenomenological couplings. So we may describe our purpose as an attempt to construct a theory in the conventional sense in which a separate field is introduced for each kind of particle. However, this is not a simple and unambiguous problem because our fundamental theory is completely  $\gamma_5$  invariant and we must make sure that this invariance is preserved at any stage of our calculations in order that the results be meaningful. For a better understanding of the problem, let us consider our Lagrangian in the lowest self-consistent approximation. We have

$$\begin{aligned} L' &= L'_0 + L'_I, \\ \text{where } L'_0 &= -(\bar{\psi}\gamma_\mu\partial_\mu\psi + m\bar{\psi}\psi), \\ L'_I &= g_0[(\bar{\psi}\psi)^2 - (\bar{\psi}\gamma_5\psi)^2] + m\bar{\psi}\psi. \end{aligned} \quad (5.1)$$

$L'$  is obviously  $\gamma_5$  invariant. In order to preserve this invariance we must study the  $S$  matrix generated by  $L'_I$ . Some subsets of diagrams have been considered in the previous section and it will be shown now how those calculations comply with  $\gamma_5$  invariance. This point must be understood clearly so that we shall discuss it in a rather systematic way. Let us recall first how we constructed the scattering matrix in the "ladder" approximation. The lowest-order contribution is certainly invariant as no internal massive line appears. But what will happen to the next-order terms [Fig. 3(b)]? To these diagrams corresponds the expression

$$J_S(q^2) - \gamma_5 J_P(q^2)\gamma_5 + iJ_{SP}(q^2)\gamma_5 + i\gamma_5 J_{PS}(q^2). \quad (5.2)$$

In the gauge in which our calculations were performed, the last two terms happened to be zero. We write down next the transformation properties of the quantities appearing above. By straightforward calculation we find

$$\begin{aligned} \gamma_5 &\rightarrow \gamma_5 \cos 2\alpha + i \sin 2\alpha, \\ 1 &\rightarrow \cos 2\alpha + i\gamma_5 \sin 2\alpha, \\ J_P &\rightarrow J_P \cos^2 2\alpha + J_S \sin^2 2\alpha, \\ J_S &\rightarrow J_S \cos^2 2\alpha + J_P \sin^2 2\alpha, \\ J_{SP} &\rightarrow (J_P - J_S) \sin 2\alpha \cos 2\alpha, \\ J_{PS} &\rightarrow (J_P - J_S) \sin 2\alpha \cos 2\alpha. \end{aligned} \quad (5.3)$$

By simple substitution the invariance follows easily. The argument can now be extended to all orders, provided at each order all the possible combinations of  $S$  and  $P$  are included. The invariance of the scattering in the "ladder" approximation is thus established. It may look surprising that the  $SP$  and  $PS$  contributions do not vanish identically. This can be understood by considering the fact that the  $\gamma_5$  transformation changes the

parity of the vacuum which will be in general a superposition of states of opposite parities. In this way products of fields of different parities (as the  $SP$  propagator) may have a nonvanishing average value in the vacuum state.

We may now attempt the construction of the phenomenological coupling by introducing two local fields  $\Phi_P$  and  $\Phi_S$  describing the pseudoscalar and the scalar particles, respectively. We start by observing that, in the same gauge in which the previous calculations were made, we can write the meson-nucleon interaction as

$$L_I = G_P i\bar{\psi}\gamma_5\psi\Phi_P + G_S i\bar{\psi}\psi\Phi_S. \quad (5.4)$$

In order to find the general expression valid in any gauge, it is convenient to introduce the following two-dimensional notation

$$\varphi = \begin{pmatrix} i\bar{\psi}\gamma_5\psi \\ \bar{\psi}\psi \end{pmatrix}, \quad \Phi = \begin{pmatrix} \Phi_P \\ \Phi_S \end{pmatrix}, \quad G = \begin{pmatrix} G_P & 0 \\ 0 & G_S \end{pmatrix}. \quad (5.5)$$

The interaction Lagrangian Eq. (5.4) can be written in this notation in a compact form,

$$L_I = \varphi G \Phi. \quad (5.6)$$

The effect of the  $\gamma_5$  transformation on  $\varphi$  is described with the aid of the matrix

$$U = \begin{pmatrix} \cos 2\alpha & -\sin 2\alpha \\ \sin 2\alpha & \cos 2\alpha \end{pmatrix}, \quad (5.7)$$

which satisfies  $UU^\dagger = UU^{-1} = UU^T = 1$ . In other words, the  $\gamma_5$  transformation induces a unitary transformation in the two-dimensional space, and Eq. (5.6) remains invariant if

$$G \rightarrow UGU^{-1}, \quad \Phi \rightarrow U\Phi. \quad (5.8)$$

To complete the construction of the theory, the free Lagrangian for the fields  $\Phi_P$  and  $\Phi_S$  must be added. If we work again in the special gauge  $\alpha=0$ , we may write

$$L_0 = -\frac{1}{2}\partial_\mu\Phi_P\partial_\mu\Phi_P - \frac{1}{2}\partial_\mu\Phi_S\partial_\mu\Phi_S - \frac{1}{2}\mu^2\Phi_S^2, \quad (5.9)$$

where  $\mu^2 = 4m^2$ . We use again the two-dimensional notation, and defining the mass operator

$$M^2 = \begin{pmatrix} 0 & 0 \\ 0 & \mu^2 \end{pmatrix}, \quad (5.10)$$

we write Eq. (5.9) in the invariant form

$$L_0 = -\frac{1}{2}\partial_\mu\Phi\partial_\mu\Phi - \frac{1}{2}\Phi M^2\Phi. \quad (5.11)$$

In this way we have given a formal prescription for the  $\gamma_5$  transformation in the phenomenological treatment. We have to emphasize here that the Lagrangians (5.9) and (5.11) are *not*  $\gamma_5$  invariant in the ordinary sense of the word. In our theory, where the mesons are only phenomenological substitutes which partially represent the dynamical contents of the theory, they may

be, however, called  $\gamma_5$  covariant. In other words, *the masses and the coupling constants are not fixed parameters, but rather dynamical quantities which are subject to transformations when the representation is changed.* It will be legitimate to ask whether this situation corresponds to the one obtained in the framework of the fundamental theory and discussed in the "ladder" approximation in the previous section. We shall examine the transformation rule for the mass operator  $M^2$ , since this illustrates the case in point. Let us calculate explicitly  $M^2$  in an arbitrary gauge  $\alpha$ . We have

$$M^2 \rightarrow UM^2U^{-1}$$

$$= \mu^2 \begin{pmatrix} \sin^2 2\alpha & -\sin 2\alpha \cos 2\alpha \\ -\sin 2\alpha \cos 2\alpha & \cos^2 2\alpha \end{pmatrix}. \quad (5.12)$$

The meaning of this equation is that the pseudoscalar and the scalar particle will have generally different masses in different gauges. In particular we see that the pseudoscalar particle has in the gauge  $\alpha$  a mass  $\sin 2\alpha \mu$ . If this is the case we must expect that after the transformation the pole in the corresponding propagator will move from  $q^2 = 0$  to  $q^2 = -(\sin^2 2\alpha) \mu^2$ . This actually may be verified directly in the "ladder" approximation which shows that the pion propagator changes according to

$$iG_F^2 \Delta_{FP} = \frac{2g_0}{1 - J_P} \rightarrow \frac{2g_0}{1 - J_P \cos^2 2\alpha - J_S \sin^2 2\alpha}. \quad (5.13)$$

Using the results of the previous section, it is seen that the denominator of the right-hand side vanishes for  $q^2 = -(\sin^2 2\alpha) 4m^2$ . In this way we have seen how our  $\gamma_5$ -invariant theory can be approximated by a phenomenological description in terms of pseudoscalar and scalar mesons. Of course one may add the vector meson as well. Such a description does not look  $\gamma_5$  invariant. It is only  $\gamma_5$  covariant, and the masses and coupling constants must be understood to be matrices which, however, can be simultaneously diagonalized.

The reason for this situation is the degeneracy of the vacuum and the world built upon it. Only after combining all the equivalent but nonintersecting worlds labeled with different  $\alpha$  do we recover complete  $\gamma_5$  invariance. Nevertheless, even in a particular world we can find manifestations of the invariance, such as the zero-mass pseudoscalar meson and the conserved  $\gamma_5$  current.

## VI. THE CONSERVATION OF AXIAL VECTOR CURRENT

In this section we will discuss another paradoxical aspect of the theory regarding the  $\gamma_5$  invariance. In Sec. 3 we argued that the  $X$  current should really be conserved, and that this is possible if a nucleon  $X$  current possesses a peculiar anomalous term. We now verify the statement explicitly in our approximation.

First we have to realize that the problem is again how to keep the  $\gamma_5$ -invariant nature of the theory at every

stage of approximation. It is well known in quantum electrodynamics that, in order to observe the ordinary gauge invariance, a certain set of graphs have to be combined together in a given approximation. The necessity for this is based on a general proof which makes use of the so-called Ward identity. In our present case there also exists an analog of the Ward identity. In order to derive it, let us first consider the proper self-energy part of our fermion in the presence of an external axial vector field  $B_\mu$  with the interaction  $L_B = -j_{\mu 5} B_\mu$ . The self-energy operator is now a matrix  $\Sigma^{(B)}(p', p)$  depending on initial and final momenta. Expanding  $\Sigma$  in powers of  $B$ , we have

$$\Sigma^{(B)}(p', p) = \Sigma(p) + \Lambda_{\mu 5}(p', p) B_\mu(p' - p) + \dots \quad (6.1)$$

We readily realize that the coefficient of the second term gives the desired  $X$ -current vertex correction.

On the other hand, the entire Lagrangian remains invariant under a local  $\gamma_5$  transformation if Eq. (2.3) is accompanied by

$$B_\mu \rightarrow B_\mu - \partial_\mu \alpha, \quad (6.2)$$

where  $\alpha$  is now an arbitrary function. In other words,

$$e^{i\alpha \gamma_5} \Sigma^{(B-\partial\alpha)} e^{i\alpha \gamma_5} = \Sigma^{(B)} \quad (6.3)$$

in a symbolical way of writing.<sup>17</sup>

Expanding (6.3) after putting  $B=0$ , we get

$$i\alpha(p' - p)[\gamma_5 \Sigma(p) + \Sigma(p') \gamma_5] = i\alpha(p' - p)(p' - p)_\mu \Lambda_{\mu 5}(p', p),$$

or

$$\gamma_5 \Sigma(p) + \Sigma(p') \gamma_5 = (p' - p)_\mu \Lambda_{\mu 5}(p', p). \quad (6.4)$$

The entire vertex  $\Gamma_{\mu 5} = i\gamma_\mu \gamma_5 + \Lambda_{\mu 5}$  then satisfies

$$\gamma_5 L'(p) + L'(p') \gamma_5 = -(p' - p)_\mu \Gamma_{\mu 5}(p', p), \quad (6.5)$$

$$L'(p) \equiv -i\gamma \cdot p - \Sigma(p),$$

which is the desired generalized Ward identity.<sup>18</sup> The right-hand side of Eq. (6.5) is the divergence of the  $X$  current, while the left-hand side vanishes when  $p$  and  $p'$  are on the mass shell of the actual particle. The  $X$ -current conservation is thus established. Moreover, the way the anomalous term arises is now clear. For if we assume  $\Sigma(p) = m$ , Eq. (6.4) gives

$$2m\gamma_5 = (p' - p)_\mu \Lambda_{\mu 5}(p' - p), \quad (6.6)$$

so that we may write the longitudinal part of  $\Lambda$  as

$$\Lambda_{\mu 5}^{(L)}(p', p) = 2m\gamma_5 q_\mu / q^2, \quad q = p' - p, \quad (6.7)$$

which is of the desired form.

Next we have to determine what types of graphs

<sup>17</sup> We assume here that  $\alpha(x)$  is different from zero only over a finite space-time region, so that the gauge of the nontrivial vacuum, which we may fix at remote past, is not affected by the transformation. The limiting process of going over to constant  $\alpha$  is then ill-defined as we can see from the fact that the anomalous term in  $\Gamma_{\mu 5}$  has no limit as  $q \rightarrow 0$ .

<sup>18</sup> See also J. Bernstein, M. Gell-Mann, and L. Michel, Nuovo cimento **16**, 560 (1960).

should be considered for  $\Gamma_\mu$  in our particular approximation of the self-energy. Examining the way in which the relation (6.3) is maintained in a perturbation expansion, we are led to the conclusion that our self-energy represented by Fig. 5(a) gives rise to the series of vertex graphs [Fig. 5(b)]. The summation of the graphs is easily carried out to give

$$\Lambda_{\mu 5} = \frac{1}{1 - J_P} J_{PA}, \quad (6.8)$$

where  $J_P$  was obtained before [Eq. (4.8)], and

$$\begin{aligned} J_{PA} &= \frac{2ig_0}{(2\pi)^4} \int \text{Tr} \gamma_5 S(p+q/2) \gamma_\mu \gamma_5 S(p-q/2) d^4 p \\ &= -\frac{g_0}{2\pi^2} i m q_\mu \int_{4m^2}^{\Lambda^2} \frac{dk^2}{q^2 + \kappa^2} (1 - 4m^2/\kappa^2)^{\frac{1}{2}}. \end{aligned} \quad (6.9)$$

Thus

$$\begin{aligned} \Gamma_{\mu 5} &= i\gamma_\mu \gamma_5 + \Lambda_{\mu 5} \\ &= i\gamma_\mu \gamma_5 + 2m\gamma^5 q_\mu / q^2, \end{aligned} \quad (6.10)$$

in agreement with the general formula. We see also that there is no form factor in this approximation.

This example will suffice to show the general procedure necessary for keeping  $\gamma_5$  invariance. When we consider further corrections, the procedure becomes more involved, but we can always find a set of graphs which are sufficient to maintain the  $X$ -current conservation. We shall come across this problem in connection with the axial vector weak interactions.

## VII. SUMMARY AND DISCUSSION

We briefly summarize the results so far obtained. Our model Hamiltonian, though very simple, has been found to produce results which strongly simulate the general characteristics of real nucleons and mesons. It is quite appealing that both the nucleon mass and the pseudo-scalar "pion" are of the same dynamical origin, and the reason behind this can be easily understood in terms of (1) classical concepts such as attraction or repulsion between particles, and (2) the  $\gamma_5$  symmetry.

According to our model, the pion is not the primary agent of strong interactions, but only a secondary effect. The primary interaction is unknown. At the present stage of the model the latter is only required to have appropriate dynamical and symmetry properties, although the nonlinear four-fermion interaction, which we actually adopted, has certain practical advantages.

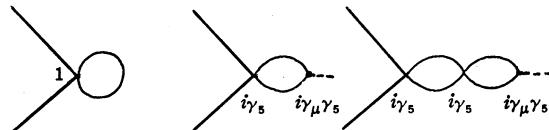


FIG. 5. Graphs for self-energy and matching radiative corrections to an axial vector vertex.



FIG. 6. A class of higher order self-energy graphs.

In our model the idealized "pion" occupies a special position in connection with the  $\gamma_5$ -gauge transformation. But there are also other massive bound states which may be called heavy mesons and deuterons. The conventional meson field theory must be regarded, from our point of view, as only a phenomenological description of events which are actually dynamic processes on a higher level of understanding, in the same sense that the phonon field is a phenomenological description of interatomic dynamics.

Our theory contains two parameters, the primary coupling constant and the cutoff, which can be translated into observed quantities: nucleon mass and the pion-nucleon coupling constant. It is interesting that the pion coupling depends only on the cutoff in our approximation. In order to make the pion coupling as big as the observed one ( $\approx 15$ ) the cutoff has to be rather small, being of the same order as the nucleon mass.

We would like to make some remarks about the higher order approximations. If the higher order corrections are small, the usual perturbation calculation will be sufficient. If they are large compared to the lowest order estimation, the self-consistent procedure must be set up, including these effects from the beginning. This is complicated by the fact that the pions and other mesons have to be properly taken into account.

To get an idea about the importance of the corrections, let us take the next order self-energy graph (Fig. 6). This is only the first term of a class of corrections shown in Fig. 6, the sum of which we know already to give rise to an important collective effect, i.e., the mesons. It would be proper, therefore, to consider the entire class put together. The correction is then equivalent to the ordinary second order self-energy due to mesons, plus modifications arising at high momenta. Thus strict perturbation with respect to the bare coupling  $g_0$  will not be an adequate procedure. Evaluating, for example, the pion contribution in a phenomenological way, we get

$$\frac{\delta m}{m} = \frac{G_P^2}{32\pi^2} \int_{m^2}^{\Lambda'^2} \frac{dk^2}{\kappa^2} \left( 1 - \frac{m^2}{\kappa^2} \right), \quad (7.1)$$

where  $\Lambda'$  is an effective cutoff. Substituting  $G_P^2$  from Eq. (4.9), this becomes

$$\frac{\delta m}{m} = \frac{1}{4} \int_{4m^2}^{4\Lambda'^2} \frac{dk^2}{\kappa^2} \left( 1 - \frac{4m^2}{\kappa^2} \right) / \int_{4m^2}^{\Lambda^2} \frac{dk^2}{\kappa^2} \left( 1 - \frac{4m^2}{\kappa^2} \right)^{\frac{1}{2}}. \quad (7.2)$$

As  $\Lambda$  and  $\Lambda'$  should be of the same order of magnitude, the higher order corrections are in general not negligible. We may point out, on the other hand, that there is a

tendency for partial cancellation between contributions from different mesons or nucleon pairs.

We already remarked before that the model treated here is not realistic enough to be compared with the actual nucleon problem. Our purpose was to show that a new possibility exists for field theory to be richer and more complex than has been hitherto envisaged, even though the mathematics is marred by the unresolved divergence problem.

In the subsequent paper we will attempt to generalize the model to allow for isospin and finite pion mass, and draw various consequences regarding strong as well as weak interactions.

#### APPENDIX

We treat here, for completeness, the problem created by the coupling of pseudoscalar and pseudovector terms encountered in the text. As we have seen, such an effect is not essential for the discussion of  $\gamma_5$  invariance, but rather adds to complication, which however naturally appears in the ladder approximation.

First let us write down the integral equation for a vertex part  $\Gamma$ :

$$\begin{aligned} \Gamma(p+\frac{1}{2}q, p-\frac{1}{2}q) &= \gamma(p+\frac{1}{2}q, p-\frac{1}{2}q) + \frac{2ig_0}{(2\pi)^4} \gamma_5 \int \text{Tr}[\gamma_5 S(p'+\frac{1}{2}q) \\ &\quad \times \Gamma(p'+\frac{1}{2}q, p-\frac{1}{2}q) S_F(p-\frac{1}{2}q)] d^4 p' \\ &- \frac{ig_0}{(2\pi)^4} \gamma_5 \gamma_\mu \int \text{Tr}[\gamma_5 \gamma_\mu S(p'+\frac{1}{2}q) \\ &\quad \times \Gamma(p'+\frac{1}{2}q, p-\frac{1}{2}q) S_F(p-\frac{1}{2}q)] d^4 p'. \end{aligned} \quad (\text{A.1})$$

This embraces three special cases depending on the inhomogeneous term  $\gamma$ :

- (a)  $\gamma = 0$  for the Bethe-Salpeter equation for the pseudoscalar meson;
- (b)  $\gamma = i\gamma_\mu \gamma_5$  for the pseudovector vertex function  $\Gamma_{\mu 5}$ ;
- (c)  $\gamma = 2g_0(\gamma_5)_f(\gamma_5)_i - g_0(\gamma_\mu \gamma_5)_f(\gamma_\mu \gamma_5)_i$  for the nucleon-antinucleon scattering through these interactions.

Here  $i$  and  $f$  refer to initial and final states, and the integral kernel of Eq. (A.1) operates on the  $f$  part.

We will consider them successively.

(a) We make the ansatz  $\Gamma = C\gamma_5 + iD\gamma_5 \gamma \cdot q$ . The integrals in Eq. (A.1) then reduce to the standard forms considered in the text. Making use of Eqs. (4.9), (4.17), and (6.9), we get<sup>16</sup>

$$\begin{aligned} C &= C - (C + 2mD)q^2 I, \\ D &= (C + 2mD)mI, \\ I(q^2) &= \frac{g_0}{4\pi^2} \int \frac{d\kappa^2}{q^2 + \kappa^2} \left(1 - \frac{4m^2}{\kappa^2}\right)^{\frac{1}{2}}, \end{aligned} \quad (\text{A.2})$$

which lead to  $q^2 = 0$ , and  $C:D = 1 - 2m^2 I(0):mI(0)$ . From Eq. (4.8), we have  $0 < 2m^2 I(0) < \frac{1}{2}$ .

$$\begin{aligned} (\text{b}) \quad \text{Put } \Gamma_{\mu 5} &= (i\gamma_\mu \gamma_5 + 2m\gamma_5 q_\mu/q^2) F_1(q^2) \\ &+ (i\gamma_\mu \gamma_5 - i\gamma \cdot q \gamma_5 q_\mu/q^2) F_2(q^2). \end{aligned} \quad (\text{A.3})$$

This is seen to satisfy the integral equation if

$$\begin{aligned} F_1 &= 1, \\ F_2 &= J_A(q^2)/[1 - J_A(q^2)], \\ J_A(q^2) &= 2m^2 I(q^2) - J_V(q^2), \end{aligned} \quad (\text{A.4})$$

where  $J(q^2)$  was defined in Eq. (4.13).

On the mass shell,  $\Gamma_{\mu 5}$  reduces to

$$\begin{aligned} &(i\gamma_\mu \gamma_5 + 2m\gamma_5 q_\mu/q^2) F(q^2), \\ F(q^2) &= 1 + F_2(q^2) = 1/[1 - J_A(q^2)]. \end{aligned} \quad (\text{A.5})$$

For  $q^2 = 0$ , we have  $J(q^2) = 0$  so that  $1 < F(0) = 1/[1 - 2m^2 I(0)] < 2$ .

(c) From the structure of the inhomogeneous term, it is clear that the scattering matrix is given by

$$M = 2g_0(\Gamma_5)_f(\gamma_5)_i + g_0(\Gamma_{\mu 5})_f(i\gamma_\mu \gamma_5)_i,$$

where  $\Gamma_5$  is the pseudoscalar vertex function.

Again, from Eq. (A.1),  $\Gamma_5$  is determined as

$$\Gamma_5 = \gamma_5 [1 - 2m^2 I(q^2)]/q^2 I(q^2) - mi\gamma \cdot q \gamma_5/q^2, \quad (\text{A.6})$$

which has an entirely different behavior from the bare  $\gamma_5$  for small  $q^2$ . The scattering matrix is then

$$\begin{aligned} M &= (\gamma_5)_f(\gamma_5)_i 2g_0 [1 - 2m^2 I(q^2)]/q^2 I(q^2) \\ &- [(i\gamma \cdot q \gamma_5)_f(\gamma_5)_i - (\gamma_5)_f(i\gamma \cdot q \gamma_5)_i] 2mg_0/q^2 \\ &- (i\gamma \cdot q \gamma_5)_f(i\gamma \cdot q \gamma_5)_i g_0 J_A(q^2)/q^2 [1 - J_A(q^2)] \\ &+ (i\gamma_\mu \gamma_5)_f(i\gamma_\mu \gamma_5)_i g_0/[1 - J_A(q^2)]. \end{aligned} \quad (\text{A.7})$$

The first three terms have a pole at  $q^2 = 0$ . The coupling constants of the pseudoscalar meson are then

pseudoscalar coupling:

$$G_p^2 = 2g_0[1 - 2m^2 I(0)]/I(0),$$

pseudovector coupling:

$$\begin{aligned} G_{pv}^2 &= g_0 J_A(0)/[1 - J_A(0)] \\ &= g_0 2m^2 I(0)/[1 - 2m^2 I(0)]. \end{aligned} \quad (\text{A.8})$$

Their relative sign is such that the equivalent pseudoscalar coupling on the mass shell is

$$G_p'^2 = 4m^2 g_0 \left\{ \left[ \frac{1 - 2m^2 I(0)}{2m^2 I(0)} \right]^{\frac{1}{2}} + \left[ \frac{2m^2 I(0)}{1 - 2m^2 I(0)} \right]^{\frac{1}{2}} \right\}^2. \quad (\text{A.9})$$

## Dynamical Model of Elementary Particles Based on an Analogy with Superconductivity. II\*

Y. NAMBU AND G. JONA-LASINIO†

*Enrico Fermi Institute for Nuclear Studies and Department of Physics, University of Chicago, Chicago, Illinois*

(Received May 10, 1961)

Continuing the program developed in a previous paper, a "superconductive" solution describing the proton-neutron doublet is obtained from a nonlinear spinor field Lagrangian. We find the pions of finite mass as nucleon-antinucleon bound states by introducing a small bare mass into the Lagrangian which otherwise possesses a certain type of the  $\gamma_5$  invariance. In addition, heavier mesons and two-nucleon bound states are obtained in the same approximation. On the basis of numerical mass relations, it is suggested that the bare nucleon field is similar to the electron-neutrino field, and further speculations are made concerning the complete description of the baryons and leptons.

### I. INTRODUCTION

**I**N Part I of this paper<sup>1</sup> we have proposed a model of strong interactions based on an analogy with the BCS-Bogoliubov theory of superconductivity. It is characterized by a nonlinear spinor field possessing  $\gamma_5$  invariance, and simulates some important features of the meson-nucleon system. The basic principle underlying the model is the idea that field theory may admit, as a result of dynamical instability, extraordinary (nontrivial) solutions that have less symmetries than are built into the Lagrangian.<sup>2</sup> In fact we have obtained as an extraordinary solution a massive fermion and a massless pseudoscalar boson as idealized proton and pion, together with other heavy mesons.

If we now try to make our model more realistic, a number of problems spring up naturally. First of all, we would have to account for the isospin and strangeness quantum numbers. It seems rather obvious that these degrees of freedom have to be built into the theory from the beginning, although there may be some possibility of utilizing both the ordinary and extraordinary solutions to enlarge the Hilbert space, as will be discussed later.

These quantum numbers will not yet be enough to determine our theory satisfactorily, as we expect to have more additional symmetries which are at least approximately satisfied. Among other things, we have postulated the  $\gamma_5$  invariance as a cornerstone of our previous model. What would be the proper generalization of the  $\gamma_5$  invariance? Then there also arises the inevitable question of any possible symmetry among baryons of different strangenesses. Since such a symmetry is at any rate only approximate, the test of the

theory will depend on its ability to account for the violation of the symmetry as well.

Finally, we face the problem of the baryon versus the lepton, the electromagnetism, and the weak processes. Here our theory creates a particular incentive for speculation concerning the baryon-lepton problem, since the ordinary and extraordinary solutions immediately remind us of these two families of particles.

We do not profess to have any clear-cut answers to these problems. In the present paper we shall again content ourselves with a rather modest task. We will first discuss a generalization of our model which incorporates the isospin for the nucleon and guarantees the existence of the pion. This can be done by demanding a  $\gamma_5 \times$  isospin gauge group with a slight violation so as to give the pion its finite mass. We find that the bare mass necessary to achieve the latter end is at most several Mev. On this basis a suggestion is made that the bare nucleon field is essentially the same as the electron-neutrino field.

The complete picture of the baryon symmetries and the baryon-lepton problem is largely beyond the scope of the present paper, but some relevant discussions on this subject will also be presented, especially those concerned with the Sakata model and the general  $\gamma_5$  symmetry.

### II. MODEL-LAGRANGIAN FOR THE NUCLEON

First we would like to observe that the nonlinear spinor field adopted in I is not an essential element of our theory, as is the case with the Heisenberg theory<sup>3</sup> but is rather a model adopted to study our dynamical principles. At least in the present stage of the game, the controlling factors are the symmetry properties and qualitative dynamical characteristics of the basic fermion-fermion interaction, and whether the interaction is due to some fundamental boson, or fundamental nonlinearity (or something entirely new) is of secondary importance. Nevertheless, we have to choose

\* This work was supported by the U. S. Atomic Energy Commission.

† Present address: Istituto di Fisica dell'Università, Roma, and Istituto Nazionale di Fisica Nucleare, Sezione di Roma, Italy.

<sup>1</sup> Y. Nambu and G. Jona-Lasinio, Phys. Rev. 122, 345 (1961); referred to hereafter as I. Y. Nambu, *Proceedings of the 1960 Annual International Conference on High-Energy Physics at Rochester* (Interscience Publishers, Inc., New York, 1960), p. 858.

<sup>2</sup> See also J. Goldstone, Nuovo cimento **19**, 154 (1961), N. N. Bogoliubov (to be published), V. G. Vaks and A. I. Larkin, *Proceedings of the 1960 Annual International Conference on High-Energy Physics at Rochester* (Interscience Publishers, Inc., New York, 1960), p. 871.

<sup>3</sup> H. P. Duerr, W. Heisenberg, H. Mitter, S. Schlieder, and K. Yamazaki, Z. Naturforsch. 14, 441 (1959); W. Heisenberg, *Proceedings of the 1960 Annual International Conference on High-Energy Physics at Rochester* (Interscience Publishers, Inc., 1960), p. 851.

some model, and naturally there will arise certain predictions specific to the particular model. We take notice of the fact that the pion, the lightest of the meson family, is pseudoscalar and isovector, whereas its isoscalar counterpart of comparable mass does not seem to exist.<sup>4</sup> If the pion is to be intimately related to a symmetry property as in our previous model, this would imply that the model of nucleons should allow an (approximate) invariance under the  $\gamma_5 \times$  isospin gauge group of Gürsey,<sup>5</sup> but not under the simple (Touschek)  $\gamma_5$  gauge group, at least not so well as in the former case. For this reason, we would altogether consider the following gauge groups:

$$\psi \rightarrow e^{i\alpha}\psi, \quad \bar{\psi} \rightarrow \bar{\psi}e^{-i\alpha}, \quad (2.1a)$$

$$\psi \rightarrow \exp(i\tau \cdot \alpha')\psi, \quad \bar{\psi} \rightarrow \bar{\psi} \exp(-i\tau \cdot \alpha'), \quad (2.1b)$$

$$\psi \rightarrow \exp(i\gamma_5\tau \cdot \alpha'')\psi, \quad \bar{\psi} \rightarrow \bar{\psi} \exp(i\gamma_5\tau \cdot \alpha''), \quad (2.1c)$$

where  $\tau$  denotes the nucleon isospin matrices.

Obviously, the first two are generators of the nucleon number gauge and the isospin transformation, respectively. The second and third transformations combined form a four-dimensional rotation group on the four components composed by the proton and neutron of both handednesses.<sup>5</sup> Thus we may also replace Eqs. (2.1a) and (2.1b) by the following transformations

$$\begin{aligned} \psi_R &\rightarrow \exp(i\tau \cdot \alpha_R)\psi_R, \quad \psi_R^\dagger \rightarrow \psi_R^\dagger \exp(-i\tau \cdot \alpha_R), \\ \psi_L &\rightarrow \exp(i\tau \cdot \alpha_L)\psi_L, \quad \psi_L^\dagger \rightarrow \psi_L^\dagger \exp(-i\tau \cdot \alpha_L), \end{aligned} \quad (2.2)$$

where  $\psi_R$  and  $\psi_L$  are the right- and left-handed components.

As the simplest Lagrangian that meets our requirements, we adopt the form

$$\begin{aligned} L = -\bar{\psi}\gamma_\mu\partial_\mu\psi - \bar{\psi}M^0\psi \\ + g_0[\bar{\psi}\psi\bar{\psi}\psi - \sum_{i=1}^3 \bar{\psi}\gamma_5\tau_i\psi\bar{\psi}\gamma_5\tau_i\psi]. \end{aligned} \quad (2.3)$$

If the bare mass operator  $M^0=0$ , this Lagrangian possesses, in addition to Eq. (2.1), an invariance under the discrete "mass reversal" group:

$$\psi \rightarrow \gamma_5\psi, \quad \bar{\psi} \rightarrow -\bar{\psi}\gamma_5. \quad (2.4)$$

The bare mass operator  $M^0$  is a possible agent for the breakdown of the Gürsey group, and will be related to the finite pion mass.<sup>6</sup> For the moment, we will assume  $M^0=0$ . Before going to solve the self-consistent equation for the mass, we give the result of the Fierz transformation on Eq. (2.3): The interaction becomes

$$\begin{aligned} L_{int} = & \frac{1}{4}g_0[\bar{\psi}\psi\bar{\psi}\psi - \bar{\psi}\gamma_5\tau_i\psi\bar{\psi}\gamma_5\tau_i\psi] \\ & + \frac{1}{4}g_0[\bar{\psi}\gamma_5\psi\bar{\psi}\gamma_5\psi - \bar{\psi}\tau_i\psi\bar{\psi}\tau_i\psi] \\ & - \frac{1}{2}g_0[\bar{\psi}\gamma_\mu\psi\bar{\psi}\gamma_\mu\psi - \bar{\psi}\gamma_\mu\gamma_5\psi\bar{\psi}\gamma_\mu\gamma_5\psi] \\ & + \frac{1}{8}g_0[\bar{\psi}\sigma_{\mu\nu}\psi\bar{\psi}\sigma_{\mu\nu}\psi - \bar{\psi}\sigma_{\mu\nu}\tau_i\psi\bar{\psi}\sigma_{\mu\nu}\tau_i\psi], \end{aligned} \quad (2.5)$$

<sup>4</sup> It may not be impossible that the ordinary  $\gamma_5$  invariance is violated more strongly than the Gürsey  $\gamma_5$  invariance so that the

which is a rather complicated combination of all kinds of terms.

We now apply the linearization procedure of I to Eqs. (2.3) and (2.4), and obtain the self-energy

$$\begin{aligned} m &= (1+\frac{1}{4})g_0 \text{Tr}S_F^{(m)}(0) \\ &= -i\frac{10g_0}{(2\pi)^4} \int \frac{d^4pm}{p^2+m^2} F(p, \Lambda). \end{aligned} \quad (2.6)$$

Note that the trace refers to both spin and isospin variables. This differs from Eq. (3.6) of I only by the change of the effective coupling  $g_0 \rightarrow 5g_0/2 \equiv g'_0$ . So we can simply take over the previous formulas, namely,

$$1 = \frac{g'_0}{4\pi^2} \int_{4m^2}^{\Lambda^2} dk^2 \left(1 - \frac{4m^2}{k^2}\right)^{\frac{1}{2}}, \quad (2.7)$$

for the nontrivial solution if the dispersion integral (4.7) of I is used.

### III. DETERMINATION OF MESON STATES

Since the interaction Lagrangian in Eqs. (1.1) and (1.3) contains a number of different couplings, we expect to get various kinds of "mesons" as bound nucleon-antinucleon pairs in our simple ladder approximation. As was explained in I, this is the proper approximation to match our self-energy equation at least for the pseudoscalar meson which is expected to have zero mass; moreover, even for other types of bound states we may reasonably trust its qualitative validity in predicting the existence and level ordering of possible bound states to the extent that our interaction is regarded basically as a short-range potential between spinor particles.

For general discussion, it is convenient to follow the procedure given in the Appendix of I. The basic equation to be considered is of the type

$$\Gamma(p + \frac{1}{2}q, p - \frac{1}{2}q) = \gamma + i \sum_n g_n O_n$$

$$\begin{aligned} &\times \int \text{Tr}[O_n S_F(p' + \frac{1}{2}q) \Gamma(p' + \frac{1}{2}q, p' - \frac{1}{2}q) \\ &\times S_F(p' - \frac{1}{2}q)] d^4p', \end{aligned} \quad (3.1)$$

where the summation on the right-hand side is over the various tensor forms in the interaction Lagrangian. The "vertex function"  $\Gamma(p + \frac{1}{2}q, p - \frac{1}{2}q)$  reduces to a bound state wave function when it becomes a homogeneous solution ( $\gamma=0$ ) for a particular value of  $q^2=-\mu^2$ . We will briefly discuss those two-nucleon states for which there is a possibility of binding.

#### A. Pseudoscalar, Isovector Meson

Unlike the case in I, only the pseudoscalar interaction  $\sim \bar{\psi}\gamma_5\tau_i\psi\bar{\psi}\gamma_5\tau_i\psi$  contributes to this state. Assuming

mass of the  $\pi_0^0$  meson may come sufficiently high. But to achieve this end by means of a bare mass does not seem to be feasible.

<sup>5</sup> F. Gürsey, Nuovo cimento **16**, 230 (1960).

<sup>6</sup> For its possible origin, see Sec. V.

$\Gamma_i^P = \gamma_5 \tau_i \Gamma^P$ , we obtain

$$\begin{aligned}\Gamma^P &= \gamma^P + \Gamma^P [1 - q^2 I^P(q^2)], \\ I^P(q^2) &= \frac{g_0'}{4\pi^2} \int_{4m^2}^{\Lambda^2} \frac{d\kappa^2}{q^2 + \kappa^2} \left(1 - \frac{4m^2}{\kappa^2}\right)^{\frac{1}{2}},\end{aligned}\quad (3.2)$$

where, of course, Eq. (2.7) was utilized. This has a homogeneous solution for  $q^2 = 0$ , corresponding to the zero-mass "pion." This pion-nucleon coupling is of pure pseudoscalar type, which can be calculated from the inhomogeneous equation with  $\gamma^P = g_0' \gamma_5 \tau_i$ , as was done in the Appendix of I. We get, namely,<sup>7</sup>

$$\begin{aligned}G_P^2/4\pi &= g_0' [4\pi I^P(0)]^{-1} \\ &= \pi \left[ \int_{4m^2}^{\Lambda^2} \frac{d\kappa^2}{\kappa^2} \left(1 - \frac{4m^2}{\kappa^2}\right)^{\frac{1}{2}} \right]^{-1}.\end{aligned}\quad (3.3)$$

### B. Scalar, Isoscalar Meson

With the ansatz  $\Gamma = \Gamma^S$  we have

$$\begin{aligned}\Gamma^S &= \gamma^S + \Gamma^S I^S(q^2), \\ I^S(q^2) &= \frac{g_0'}{4\pi^2} \int_{4m^2}^{\Lambda^2} \frac{\kappa^2 - 4m^2}{q^2 + \kappa^2} d\kappa^2 \left(1 - \frac{4m^2}{\kappa^2}\right)^{\frac{1}{2}} \\ &= 1 - (q^2 + 4m^2) I^P(q^2).\end{aligned}\quad (3.4)$$

This leads to a zero-binding state:  $q^2 = -4m^2$  with the scalar nucleon coupling constant

$$G_S^2/4\pi = g_0' [4\pi I^P(-4m^2)]^{-1}.\quad (3.5)$$

### C. Vector Mesons

There are two vector mesons, with isospin 1 and 0. The isovector meson arises from the tensor interaction  $\sim \bar{\psi} \sigma_{\mu\nu} \tau_i \psi \bar{\psi} \sigma_{\mu\nu} \tau_i \psi$  with the wave function of the type

$$\Gamma_{\mu i}^V = \sigma_{\mu\nu} q_\nu \tau_i \Gamma^V.$$

The mass is determined from<sup>8</sup>

$$\begin{aligned}1 &= -\frac{g_0'}{60\pi^2} \int_{4m^2}^{\Lambda^2} \frac{d\kappa^2}{\kappa^2 - \mu^2} \left(1 - \frac{4m^2}{\kappa^2}\right)^{\frac{1}{2}} \\ &\quad \times \left[ \kappa^2 - 4m^2 - \mu^2 \left(2 + \frac{4m^2}{\kappa^2}\right) \right],\end{aligned}$$

which has a solution (for sufficiently small  $\Lambda^2$ )

$$\mu^2 \geq 10m^2/3.$$

The coupling of this meson to the nucleon is necessarily of the derivative type.

<sup>7</sup> Note that this is half the value of  $I$  because a pion (e.g.,  $\pi_0$ ) consists of two substates  $\bar{p}p$  and  $\bar{n}n$ , which changes the normalization of the pion wave function.

<sup>8</sup> The ambiguity about the subtraction of the most divergent part was discussed in I, section 4. The gross qualitative feature is not altered even if we do not make a subtraction.

To the isoscalar meson, both vector and tensor interactions contribute, the former being attractive and the latter repulsive. The wave function will have the form

$$\Gamma_\mu^V = \gamma_\mu \Gamma_1^V + \sigma_{\mu\nu} q_\nu \Gamma_2^V,$$

which yields a coupled equation for  $\Gamma_1$  and  $\Gamma_2$ . This coupling, however, is rather small, so that we get a solution by neglecting  $\Gamma_2$ :

$$\begin{aligned}1 &= \frac{g_0'}{15\pi^2} \mu^2 \int_{4m^2}^{\Lambda^2} \frac{d\kappa^2}{\kappa^2 - \mu^2} \left(1 - \frac{4m^2}{\kappa^2}\right)^{\frac{1}{2}} \left(1 + \frac{2m^2}{\kappa^2}\right), \\ \mu^2 &\geq 20m^2/7.\end{aligned}$$

The nuclear coupling will be predominantly non-derivative.

### D. The "Deuteron" States

As in I, we can discuss the nucleon-nucleon states in parallel with the meson states. The interaction may be written conveniently in the form

$$L_{\text{int}} = \frac{1}{4} g_0 [\bar{\psi} \gamma_\mu \psi^c \bar{\psi}^c \gamma_\mu \psi - \bar{\psi} \sigma_{\mu\nu} \psi^c \bar{\psi}^c \sigma_{\mu\nu} \psi + \bar{\psi} \gamma_\mu \gamma_5 \tau_i \psi^c \bar{\psi}^c \gamma_\mu \gamma_5 \tau_i \psi].$$

This is seen to lead to two bound states: a pseudovector, isoscalar ( $J=1^+$ ,  $T=0$ ) coming from the first two interaction terms, and a scalar, isovector ( $J=0^+$ ,  $T=1$ ), coming from the last term. For the  $J=1^+$ ,  $T=0$  state (deuteron) the main contribution comes from the attractive tensor interaction, and we get

$$\begin{aligned}\Gamma_\mu &= \sigma_{\mu\nu} q_\nu \Gamma^{A'}, \\ 1 &= \frac{g_0'}{4\pi^2} \int_{4m^2}^{\Lambda^2} \frac{d\kappa^2}{\kappa^2 - \mu^2} \left(1 - \frac{4m^2}{\kappa^2}\right)^{\frac{1}{2}} \left[\mu^2 + \frac{2}{15}(\kappa^2 - 4m^2)\right], \\ \mu^2 &\geq 17m^2/5.\end{aligned}$$

For the  $J=0^+$ ,  $T=1$  case we have

$$\begin{aligned}\Gamma &= \gamma_5 \gamma \cdot q \Gamma^{S'}, \\ 1 &= \frac{4}{5} m^2 I^P(-\mu^2), \\ \mu^2 &\geq 16m^2/5.\end{aligned}$$

### IV. VIOLATION OF $\gamma_5$ INVARIANCE

Let us now discuss the violation of the  $\gamma_5$  invariance as indicated by the finite mass of the real pion. It would be senseless, of course, to talk about the invariance if the observed pion mass implied a large departure from our original Lagrangian, for example, due to a bare nucleon mass as large as the observed mass. So we need to estimate the amount of violation in the Lagrangian.

In general, the bare mass operator, which does not

violate nucleon number conservation,<sup>9</sup> can have the following form

$$M^0 = m_1^0 + m_2^0 \boldsymbol{\tau} \cdot \mathbf{n} + m_3^0 i\gamma_5 + m_4^0 i\gamma_5 \boldsymbol{\tau} \cdot \mathbf{n}', \quad (4.1)$$

where  $\mathbf{n}$  and  $\mathbf{n}'$  are arbitrary unit vectors in the isospin space. The observed mass generated by Eq. (4.1) will also have a similar form. Because of the invariance of the rest of the Lagrangian under the transformations (2.1), we can choose it to be

$$M = m_1 + m_2 \tau_3 + m_3 i\gamma_5, \quad (4.2)$$

which gives two eigenmasses

$$\begin{aligned} m_p &\equiv [(m_1 + m_2)^2 + m_3^2]^{\frac{1}{2}}, \\ m_n &\equiv [(m_1 - m_2)^2 + m_3^2]^{\frac{1}{2}}. \end{aligned} \quad (4.3)$$

The self-consistent self-energy equation to be solved is now

$$\begin{aligned} M = M^0 + g_0 \{ & \text{Tr} S_F^{(M)}(0) - \gamma_5 \tau_i \text{Tr} [\gamma_5 \tau_i S_F^{(M)}(0)] \\ & + \frac{1}{5} \gamma_5 \text{Tr} [\gamma_5 S_F^{(M)}(0)] - \frac{1}{5} \tau_i \text{Tr} [\tau_i S_F^{(M)}(0)] \}. \end{aligned} \quad (4.4)$$

Equating the respective coefficients of both sides, we get

$$\begin{aligned} m_1 &= m_1^0 + \bar{I}m_1 + \bar{I}m_2, \\ m_2 &= m_2^0 - \frac{1}{5}\bar{I}m_2 - \frac{1}{5}\bar{I}m_1, \\ m_3 &= m_3^0 - \frac{1}{5}\bar{I}m_3, \\ 0 &= m_4^0 + \bar{I}m_3, \end{aligned} \quad (4.5)$$

where

$$\begin{aligned} \bar{I} &= \frac{1}{2}[I(m_p) + I(m_n)], \\ \bar{I}' &= \frac{1}{2}[I(m_p) - I(m_n)], \\ I(m) &= -\frac{8ig_0'}{(2\pi)^4} \int \frac{d^4 p}{p^2 + m^2} F(p, \Lambda). \end{aligned} \quad (4.5')$$

We are interested in a small change of the non-trivial solution due to  $M^0$ . From Eq. (4.5) it is clear that  $m_3 = 0$  unless

$$m_3^0 = -(\frac{1}{5} + \bar{I}^{-1})m_4^0 \neq 0.$$

The term  $m_3$  implies a violation of time and space reflections. Since we are not interested in such a violation, we will assume  $m_3 = m_3^0 = m_4^0 = 0$  from now on. We further note that

$$\bar{I} = I(m_1) + O[(m_2/m_1)^2], \quad \bar{I}' = O[m_2/m_1].$$

In fact, up to the first order in  $m_2/m_1$ , we may put

$$m_1 = m_1^0 + I(m_1)m_1, \quad (4.6a)$$

$$m_2 = m_2^0 - \frac{1}{5}[I(m_1) + 2I'(m_1)m_1^2]m_2, \quad (4.6b)$$

<sup>9</sup> The most general form of the self-energy Lagrangian (neglecting isospin dependence) is

$$\begin{aligned} \psi [i\gamma \cdot p \Sigma_1(p^2) + \Sigma_2(p^2) + i\gamma \cdot p \gamma_5 \Sigma_3(p^2) + i\gamma_5 \Sigma_4(p^2)]\psi \\ + \bar{\psi} [i\gamma \cdot p \Sigma_5(p^2) + \Sigma_6(p^2) + i\gamma \cdot p \gamma_5 \Sigma_7(p^2) + i\gamma_5 \Sigma_8(p^2)]\bar{\psi} \\ + H.c. \end{aligned}$$

We do not attempt to study such a problem at this place.

where

$$I'(m) = dI(m)/d(m^2) < 0.$$

Equation (4.6a) determines  $m_1$  in terms of  $m_1^0$ .

The self-consistency condition required, for  $m_1^0 = 0$ , is that  $I(m) = 1$ . We may thus expand  $I(m)$ :

$$I(m_1) = 1 + I'(m)(m_1^2 - m^2),$$

and obtain

$$\Delta m^2 = m_1^2 - m^2 = -m_1^0 / [mI'(m)]. \quad (4.7)$$

Since  $I'(m)$  is of the order of  $-I(m)/m^2$  (see below), this means

$$\Delta m = m_1 - m \approx m_1^0. \quad (4.8)$$

From (4.6b), then

$$\begin{aligned} m_2 &\approx m_2^0 \{1 - \frac{1}{5}[1 + I'(m)m^2]\}^{-1} \\ &\approx m_2^0. \end{aligned} \quad (4.8')$$

We note that originally there were two solutions  $\pm |m|$ , which now split into opposite directions according to Eq. (4.7) or (4.8). The meaning of this is as follows. Under the strict  $\gamma_5$  invariance, there is a complete degeneracy with respect to the transformation (2.1c). The perturbation  $m_1^0$  removes this degeneracy, so that the energy of the vacuum will depend on the orientation of the “ $\gamma_5$  spin” of the negative energy fermions present in the “vacuum” with respect to this preferred direction. Obviously, the self-consistent procedure, which is similar to the variational method, gives the two extremum configurations corresponding to parallel ( $m_0/m > 0$ ) or antiparallel ( $m_0/m < 0$ )  $\gamma_5$ -spin lineup. The parallel case has the larger “gap parameter”  $|m|$  than the antiparallel case, so that the former will correspond to the stable ground state. The latter, on the other hand, should correspond to a metastable world.

It is perhaps interesting to see the general behavior of the self-consistency equation for arbitrary magnitude of  $m_1^0$ , assuming  $m_2^0 = 0$  for simplicity. The relevant equation,

$$m[1 - I(m)] = m^0,$$

is plotted schematically in Fig. 1.

Note that the trivial branch of the solution, which goes through the origin, has  $m_0/m < 0$ . In other words, even in this case the self-consistent solution is qualitatively different from the simple perturbation result. As  $m^0$  increases, it approaches the metastable nontrivial solution, and finally both go into the complex plane.

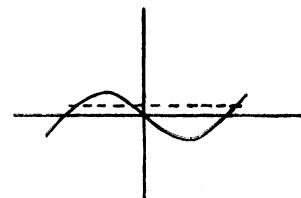


FIG. 1. The three self-consistent mass solutions  $m$  (ordinate) as a function of the bare mass  $m^0$  (abscissa).

We now come to the meson problem. The pion mass will be determined from

$$\Gamma_j = ig_0' \tau_i \int \text{Tr}[\tau_i S_F(p' + \frac{1}{2}q) \Gamma_j S_F(p' - \frac{1}{2}q)] d^4 p', \quad (4.9)$$

but

$$\begin{aligned} & \underset{\text{isospin}}{\text{Tr}} [\tau_i S_F^{(M)} \tau_j S_F^{(M)}] \\ &= \underset{\text{isospin}}{\text{Tr}} \left[ \tau_i \left( S_F^{(m_p)} \frac{1+\tau_3}{2} + S_F^{(m_n)} \frac{1-\tau_3}{2} \right) \right. \\ & \quad \times \left. \tau_j \left( S_F^{(m_p)} \frac{1+\tau_3}{2} + S_F^{(m_n)} \frac{1-\tau_3}{2} \right) \right] \\ &= 2\delta_{ij} \frac{S_F^{(m_p)} + S_F^{(m_n)}}{2} \frac{S_F^{(m_p)} + S_F^{(m_n)}}{2} \\ & \quad + 2(2\delta_{ij}\delta_{j3} - \delta_{ij}) \frac{S_F^{(m_p)} - S_F^{(m_n)}}{2} \frac{S_F^{(m_p)} - S_F^{(m_n)}}{2}. \end{aligned}$$

The second term yields convergent results, and is  $O[(\Delta m/m)^2]$ . To the order  $\Delta m/m$ , therefore, only the first term is important; moreover,

$$(S_F^{(m_p)} + S_F^{(m_n)})/2 \approx S_F^{(m_1)}.$$

In other words, there will be no first-order mass splitting of the pion. The mass is then determined from

$$\begin{aligned} 1 &= J_{p1}(-\mu^2) \\ &= \frac{g_0'}{4\pi^2} \int_{4m_1^2}^{\Lambda^2} \frac{\kappa^2 d\kappa^2}{\kappa^2 - \mu^2} \left( 1 - \frac{4m_1^2}{\kappa^2} \right)^{\frac{1}{2}}. \end{aligned} \quad (4.10)$$

For  $m_1^0 = 0$ , we had originally

$$1 = J_p(0) = \frac{g_0'}{4\pi^2} \int_{4m_1^2}^{\Lambda^2} d\kappa^2 \left( 1 - \frac{4m_1^2}{\kappa^2} \right)^{\frac{1}{2}},$$

which should now be replaced by

$$1 = \frac{m_1^0}{m_1} + \frac{g_0'}{4\pi^2} \int_{4m_1^2}^{\Lambda^2} d\kappa^2 \left( 1 - \frac{4m_1^2}{\kappa^2} \right)^{\frac{1}{2}}, \quad (4.11)$$

according to Eq. (4.6a).

From Eqs. (4.10) and (4.11) follows

$$\begin{aligned} \frac{m_1^0}{m_1} &= \mu^2 \frac{g_0'}{4\pi^2} \int_{4m_1^2}^{\Lambda^2} \frac{d\kappa^2}{\kappa^2 - \mu^2} \left( 1 - \frac{4m_1^2}{\kappa^2} \right)^{\frac{1}{2}} \\ &\approx \mu^2 \frac{g_0'}{4\pi^2} \int_{4m_1^2}^{\Lambda^2} \frac{d\kappa^2}{\kappa^2} \left( 1 - \frac{4m_1^2}{\kappa^2} \right)^{\frac{1}{2}} \\ &\leq \frac{\mu^2}{4m_1^2} \left( 1 - \frac{m_1^0}{m_1} \right). \end{aligned} \quad (4.12)$$

For the observed value of  $\mu^2/4m_1^2 \approx 1/200$  we then have, for the stable solution

$$m_1^0 \lesssim m_1/200 \approx 5 \text{ Mev.} \quad (4.13)$$

The amount of bare mass needed to produce the pion mass is thus surprisingly small.

On the other hand, the metastable solution ( $m_1^0/m < 0$ ) produces an imaginary pion mass, indicating the unphysical nature of the solution.

The pion-nucleon coupling constant at the pion pole becomes [see Eq. (2.3)]

$$G_P^2/4\pi \approx g_0' [4\pi I_P(-\mu^2)]^{-1},$$

which is changed from the old one only by an order  $\mu^2/m_1^2 \sim m_1^0/m_1$ .

The other heavy meson states can be treated similarly. We see easily that the changes induced by  $m_1^0$  are quite small: In general  $\Delta\mu^2/m_1^2 = O(m_1^0/m_1)$  and  $\Delta G^2/G^2 = O(m_1^0/m_1)$ . Thus the effect of  $M^0$  shows up dramatically only in the pion mass because it was originally zero.

Finally we remark that instead of a bare mass, we could assume slightly different coupling constants  $g_s$  and  $g_p$  ( $< g_s$ ) for the scalar and pseudoscalar interaction terms in the Lagrangian (2.3). The nature of the solution is somewhat different from the previous case because the Lagrangian still retains the mass reversal invariance  $\psi \rightarrow \gamma_5 \psi$ , and the solution is twofold degenerate ( $\pm m$ ). The fractional change of the coupling necessary to produce the pion mass is again small:  $|\Delta g/g| \approx \mu^2/4m_1^2$ .

## V. IMPLICATIONS OF THE MODEL

Let us now discuss the relevance of our present model to the physical realities of the nucleons and mesons.

1. We have seen that our Lagrangian (2.3) leads to the nucleon of isospin  $\frac{1}{2}$  and the pion of isospin 1. The pion-nucleon coupling constant (pseudoscalar) depends on the cutoff parameter. For the observed large value ( $\approx 15$ ) of  $G_P^2/4\pi$ , we see from Eqs. (1.5) and (2.3) that  $\Lambda$  must be of the same order of magnitude as the nucleon mass itself. This is not unreasonable, since the effective nucleon-nucleon interaction in higher approximations would proceed with the exchange of nucleon pairs.

A third parameter, the bare mass, enters our picture in order to make the meson mass finite. It would seem rather unsatisfactory and embarrassing that after all one has to break the postulated symmetry in an *ad hoc* manner. In order to clear up this point, the origin of the effective bare mass then becomes an interesting and important question. Since the required bare mass [Eq. (4.13)] seems to be quite small, a tempting possibility suggests itself that the bare nucleon field is the same as the electron-neutrino field. The electron mass itself could be either intrinsic or of electromagnetic

origin.<sup>10</sup> Under this assumption, the bare mass operator would have the form  $M^0 = m_e^0(1 + \tau_3)/2$ , where the word "bare" is used relative to the interaction under consideration. According to the results of the previous section, it is only the isoscalar part of  $M^0$  that produces the large shift of the pion mass, and the amount of violation of the isospin invariance will remain small.

2. Besides the pion, we have also derived vector mesons of both isoscalar ( $T=0$ ) and isovector ( $T=1$ ) types, and a scalar isoscalar meson which is actually unbound. No state corresponding to the isoscalar pion ( $\pi^0$ ) is found. Of course, these results should depend sensitively on the choice of the interaction in the first place, and to a lesser extent also on the degree of approximation. At any rate, it seems to be a rather interesting and satisfactory feature of the model that these same vector mesons have been anticipated theoretically from various grounds,<sup>11</sup> even though there do not seem to be convincing experimental indications of their existence as yet.<sup>12</sup>

The mass values obtained here are rather high, and these mesons should actually decay into pions very quickly. The coupling constants are generally of the same order as the pion coupling constant, which means a very strong interaction for the vector and scalar mesons. These results, however, may be considerably altered in a better approximation. For one thing, the heavy mesons are coupled strongly to many-pion states which would make the former mere resonances of the latter. Moreover, the nucleon-nucleon and meson-meson interactions can go through long-range forces due to the exchange of these same mesons, which would in turn change the meson states themselves. These processes (the so-called left-hand cuts in the language of the dispersion theory) have not been taken into account in our ladder approximation.

This is a highly cooperative mechanism, and if one wants to handle it in a systematic way, one may be led to the same dispersion theoretical approach that is now widely pursued in pion physics. As a result of such effects, it is then conceivable that the masses of the vector mesons, for example, may come down.<sup>13</sup> Al-

<sup>10</sup> The electromagnetic interaction is invariant under the simple  $\gamma_5$  transformation, but not under the Gürsey transformation since it fundamentally distinguishes between the charged and neutral components. Thus there is a built-in violation which can eventually produce the pion mass.

<sup>11</sup> W. R. Frazer and J. R. Fulco, Phys. Rev. **117**, 1609 (1960); Y. Nambu, *ibid.* **106**, 1366 (1957); G. Chew, Phys. Rev. Letters **4**, 142 (1960); J. J. Sakurai, Ann. Phys. **11**, 1 (1960).

<sup>12</sup> J. A. Anderson, Vo X. Bang, P. G. Burke, D. D. Carmony, and N. Schmitz, Phys. Rev. Letters **6**, 365 (1961); A. Abashian, N. Booth, and K. M. Crowe, *ibid.* **5**, 258 (1960).

<sup>13</sup> A crude way to see the general tendency will be to argue as follows: The  $T=1$  vector meson is coupled to the nucleon mainly through tensor coupling, so that it will cause a nucleon-antinucleon interaction of the type  $-g^2(\sigma_1 \cdot \sigma_2 \tau_1 \cdot \tau_2)e^{-\mu r}/r$ . This tends to raise  $T=1, J=0^+$  and  $T=0, J=1^-$  meson states, and

ternatively, it is also conceivable that we have more than one resonance having the same quantum numbers, of which we have obtained the higher ones. These high-energy poles may in turn determine the low-energy resonances.

In addition to the vector mesons, we expect a  $T=0, J=0^+$  resonance, which has also been postulated by some people.<sup>14</sup> We should try to check these predictions against experimental evidence, such as the characteristic  $Q$ -value distributions and angular correlations in meson production processes.

Turning to the nucleon number 2 states, we expect two bound states ( $T=0, J=1^+$  and  $T=1, J=0^+$ ) with comparable masses to those for the vector mesons. This is a qualitatively satisfactory feature in view of the observed deuteron and the singlet virtual states, even though the actual binding is considerably weaker.<sup>15</sup>

3. As was already mentioned in I, our particular model was motivated by the approximate axial vector conservation observed in the nuclear  $\beta$  decay and the role of the pion in it.<sup>5,16</sup> The only difference from I is that (a) we now have the conservation of the isovector axial vector current  $i\bar{\psi}\gamma_\mu\gamma_5\tau_3\psi$  instead of the simple axial vector current  $i\bar{\psi}\gamma_\mu\gamma_5\psi$ , and (b) a small violation of conservation is explicitly introduced. The general treatment of the problem will be completely analogous to the previous case.

Assuming that the  $\beta$  decay occurs through an additional term in the Lagrangian

$$L_\beta = g_\beta \bar{\psi} \gamma_\mu (1 + \gamma_5) \tau_3 \psi l_\mu + \text{H.c.} \quad [\tau_+ = \frac{1}{2}(\tau_1 + i\tau_2)],$$

where  $l_\mu$  refers to the lepton current, the nuclear  $\beta$ -

lower  $T=0, J=0^+$  and  $T=1, J=1^-$  states. Any change in the binding force, however, will be offset by the corresponding change in the nucleon mass, which automatically adjusts the pion mass to lie where it should be. The exchange of the  $T=0, J=1$ , and  $J=0$  mesons, therefore, would not be so important in determining the relative shift of the meson levels.

<sup>14</sup> J. Schwinger, Ann. Phys. **2**, 407 (1957); M. Gell-Mann and M. Lévy, Nuovo cimento **16**, 705 (1960); S. Gupta, Phys. Rev. **111**, 1436 (1958), Phys. Rev. Letters **2**, 124 (1959); M. H. Johnson and E. Teller, Phys. Rev. **98**, 783 (1955); H. P. Duerr and E. Teller, *ibid.* **103**, 469 (1956). The  $\sigma$  meson mass obtained here is independent of the cutoff  $\Lambda$ , so that there may be some point in arguing that it is more reliable than for the vector mesons. If so, we may expect a nucleon-antinucleon resonance near zero kinetic energy (taking account of the mass shift due to  $M^0$ ). The width may be quite broad.

<sup>15</sup> In fact, both  $T=0$  and  $T=1$  vector meson exchanges work in the direction to reduce the binding relative to the nucleon-antinucleon case.

<sup>16</sup> S. Bludman, Nuovo cimento **9**, 433 (1958); F. Gürsey, Ann. Phys. **12**, 91 (1961); Y. Nambu, reference 1; M. Gell-Mann and M. Levy, reference 14; J. Bernstein, N. Gell-Mann, and L. Michel, Nuovo cimento **16**, 560 (1960); J. Bernstein, S. Fubini, M. Gell-Mann, and W. Thirring, *ibid.* **17**, 757 (1960); Chou Kuang-Chao, J. Exptl. Theoret. Phys. (U.S.S.R.) **39**, 703 (1960) [Soviet Phys.—JETP **12**, 492 (1961)].

decay vertex becomes

$$\Gamma_\mu = g_\beta [i\gamma_\mu \tau_+ F_{V1}(q^2) - i\sigma_{\mu\nu} q_\nu \tau_+ F_{V2}(q^2) + \{i\gamma_\mu \gamma_5 \tau_+ \\ + [2m_1 \gamma_5 q_\mu \tau_+ / (q^2 + \mu_\pi^2)] f(q^2)\} F_A(q^2)],$$

where  $q$  is the momentum change. In the ladder approximation,  $F_{V1}(q^2)$  arises from the vector-type nucleon pairs, and  $F_{V1}(0)=1$  (in accordance with the Ward identity, applicable to the isospin current, which shows that  $F_{V1}(0)=1$  in general.<sup>17</sup>)

In the axial vector part,  $F_A(q^2)=1$  in our approximation.  $f(q^2)$  arises because of the violation of the  $\gamma_5$  invariance, but it deviates from 1 only to the order  $m_1^0/m_1 \sim \mu^2/m_1^2$ , as was already seen in the previous section. For practical purposes, therefore, the axial vector current has the desired form which would lead to the Goldberger-Treiman relation<sup>18</sup>

$$2m_1 g_A \approx \sqrt{2} G_\pi g_\pi,$$

where  $g_A = g_\beta F_A(0)$  and  $G_\pi$ ,  $g_\pi$  are, respectively, pion-nucleon and pion-lepton couplings.

In higher orders, however,  $F_A(q^2)$  will be present, and in general  $F_A(0) \neq 1$  even under the strict  $\gamma_5$  invariance. People have conjectured in the past that  $F_A(0) = g_A/g_V = 1$  as  $\mu_\pi \rightarrow 0$ , but this does not seem to be easily guaranteed. The generalized Ward identity for the axial vector current<sup>19</sup> suffices to prove the Goldberger-Treiman relation, but is not enough to make  $F_A(0)=1$ . In order that the latter should come out rigorously, we would need a more subtle mechanism. Nevertheless, we can try a working hypothesis that  $g_A/g_V=1$  under the strict invariance, and then estimate the deviation due to the violation. This scheme is carried out in the Appendix.

## VI. FURTHER PROBLEM

We will consider here some of the general problems which have not been explored, but which seem to be important in a more comprehensive understanding of the elementary particles.

*1. The hyperons.* In order to incorporate the strange particles into our picture we would have to increase the dimensions of the fundamental field unless we do further unconventional things (see below). The simplest possibility from the point of view of quantum numbers would be to add a bare  $\Lambda$ -particle field as was originally proposed by Sakata.<sup>20</sup> We would then postulate, in addition, the generalized  $\gamma_5$  symmetry, which would mean the invariance of the left-handed and right-handed components separately under the unitary transformation among the three fields or some subgroups of it. The mass splitting of the three baryons will be obtained

<sup>17</sup> R. P. Feynman and M. Gell-Mann, Phys. Rev. **109**, 193 (1958).

<sup>18</sup> M. L. Goldberger and S. B. Treiman, Phys. Rev. **111**, 356 (1958).

<sup>19</sup> J. Bernstein *et al.*, reference 16.

<sup>20</sup> S. Sakata, Progr. Theoret. Phys. (Kyoto) **16**, 686 (1956).

from bare masses of similar magnitude, which destroys the otherwise rigorous symmetry.

This approach will produce easily the pions and  $K$  mesons and probably more, and their masses can again be related to the baryon bare masses. But we do not yet have a comparable dynamical method to predict  $\Sigma$  and  $\Xi$  particles. Consequently, we shall not be able to say whether or not the present model is dynamically satisfactory in this respect.

*2. The leptons.* In connection with the above model we are naturally led to the lepton problem. Gamba, Marshak, and Okubo<sup>21</sup> have pointed out an interesting parallelism between the  $p\pi\Lambda$  and  $\nu e\mu$  triplets. As was remarked in the beginning, our theory gives a special incentive for speculation about this relation because we have obtained two solutions: one ordinary and one extraordinary, differing in masses. Could they both be realized in nature simultaneously? According to our results in I, the answer is no because they belong to different Hilbert spaces. Moreover, the trivial solution gives rise to unphysical mesons at least under the assumption of fixed cutoff, with a large mass ( $-\mu^2 \gtrsim \Lambda^2$ ) but not necessarily a weak coupling ( $G^2 \lesssim \Lambda^4/\mu^4$ ). Nevertheless, it would seem too bad if Nature did not take advantage of the two solutions. A straightforward way to make the two solutions co-exist in the same world is obviously to postulate that the world is represented by the direct product of two Hilbert spaces<sup>22</sup>:

$$\mathcal{H} = \mathcal{H}^{(0)} \otimes \mathcal{H}^{(m)}, \quad (6.1)$$

built upon the vacuum state

$$\Omega = \Omega^{(0)} \otimes \Omega^{(m)}. \quad (6.1')$$

It is true that this is effectively the same as doubling the fields, but here the choice of the two solutions (particles) is dictated by the dynamics of the original nonlinear theory. In order to describe this situation, we may adopt an *effective* Lagrangian

$$L = L^{(1)} + L^{(2)}, \quad (6.2)$$

where each of the  $L^{(i)}$  has the same form, only differing in the charge assignments of the respective triplet fields. The Lagrangian obviously yields four subspaces

$$\mathcal{H}_1^{(0)} \otimes \mathcal{H}_2^{(m)}, \quad \mathcal{H}_1^{(m)} \otimes \mathcal{H}_2^{(0)}, \quad \mathcal{H}_1^{(0)} \otimes \mathcal{H}_2^{(0)}$$

and

$$\mathcal{H}_1^{(m)} \otimes \mathcal{H}_2^{(m)}.$$

According to our plan, we must say that we happen to live in the first subspace. [In the second space, the masses of  $\nu e\mu$  and  $p\pi\Lambda$  are interchanged, whereas in the third (fourth) case we have two kinds of leptons (baryons).]

<sup>21</sup> A. Gamba, R. E. Marshak, and S. Okubo, Proc. Natl. Acad. Sci. U. S. **45**, 881 (1959); Z. Maki, M. Nakagawa, Y. Ohnuki, and S. Sakata, Progr. Theoret. Phys. **23**, 1174 (1960).

<sup>22</sup> S. Okubo and R. E. Marshak, Nuovo cimento **19**, 1226 (1961), have independently proposed a similar idea. We thank the authors for valuable communications.

So far there is no interaction between leptons and baryons (except the electromagnetic, which is trivial). To introduce the weak interactions, we may, for example, add to Eq. (5.2) a third nonlinear term involving all the (left-handed) fields. This would complete our program of dealing with the strong and weak interactions.

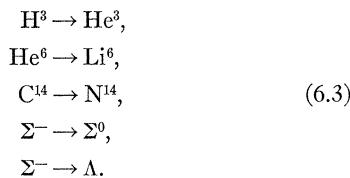
But, of course, it is not yet a truly unified theory; the weak interaction is introduced only as an *ad hoc* additional process. Moreover, we do not know the mathematical consistency of such a procedure, because the additional interaction, if taken seriously, may qualitatively affect the baryon and lepton solutions we already have.

There is an alternative, but less drastic scheme; namely, to assume six different fields from the beginning, of which three (becoming eventually the baryon fields) have additional strong interactions in the Lagrangian. This may not be devoid of elegance if the interaction is mediated by a vector Bose field coupled to the baryon charge. The intermediate bosons, including the photons and possibly also the weak bosons, could then be interpreted as the agents that distinguish between different components of the bare fermions, which otherwise would enjoy a high degree of symmetry.

We would like to throw in another remark here that there may be also a possibility of utilizing the ordinary and extraordinary solutions in distinguishing between electron and muon, or baryons of different strangenesses.

3. *The  $\gamma_5$  invariance for general systems.* In our theory the  $\gamma_5$  invariance is a very essential element. It is a particular symmetry which exists in the Lagrangian, but is masked in reality because of the (approximate) degeneracy of the vacuum with respect to that symmetry. We have used the pion and the  $\beta$  decay in support of the assumption. In order to firmly establish its validity, however, we must try to find more evidences. For one thing, the induced pseudoscalar terms in nucleon  $\beta$  decay and  $\mu$  capture should be examined more closely.

Furthermore, if such a symmetry is to have a general meaning, we must be able to consider partially conserved currents for processes such as



An elementary definition of the  $\gamma_5$  transformation for the general system is obvious: When the wave function of a system is expressed in terms of the fundamental (bare) spinors obeying the rules Eq. (2.1), the transformation is unambiguously defined for each com-

ponent, and thereby the total axial vector current is determined.

For superallowed transitions with spin  $\frac{1}{2}$ , the problem is particularly simple, since it is the same as for the neutron case. Thus for  $\text{H}^3 \rightarrow \text{He}^3$  we have the same Goldberger-Treiman relation

$$(M_{\text{H}} + M_{\text{He}})g_A(\text{H}, \text{He})/\sqrt{2}G_\pi(\text{H}, \text{He}) \approx g_\pi, \quad (6.4)$$

where  $g_A$ ,  $G_\pi$  now characterize the  $\beta$  decay and the (unknown) pion coupling for the transition under consideration.

Similar relations hold for the  $\Sigma$  decays.<sup>23</sup> For the  $\Sigma-\Sigma$  case, we have

$$2m_\Sigma g_A(\Sigma\Sigma)/G_\pi(\Sigma\Sigma) \approx g_\pi. \quad (6.5)$$

For the  $\Sigma-\Lambda$  case, the axial vector vertex becomes<sup>24</sup>

$$\begin{aligned} \Gamma_A \approx & [i\gamma_\mu\gamma_5 + (m_\Sigma + m_\Lambda)\gamma_5 q_\mu/(q^2 + \mu_\pi^2)]F_{1A}(q^2) \\ & + i\gamma_5 \sigma_{\mu\nu} q_\nu F_{2A}(q^2) \\ (m_\Sigma + m_\Lambda)F_{1A}(0)/G_\pi(\Sigma\Lambda) \approx & g_\pi, \end{aligned} \quad (6.6)$$

if the relative  $\Sigma-\Lambda$  parity is even. The vector current conservation is also violated because of the  $\Sigma-\Lambda$  mass difference, and it looks as though this would predict a corresponding scalar meson term. However, the analogy is rather superficial. Firstly, the violation disappears if  $m_\Sigma = m_\Lambda$ , in which case there would be no need for a scalar meson. The  $\Sigma-\Lambda$  mass difference itself might be due to the breakdown of the  $\gamma_5$  symmetry. Secondly, it is an "unfavored" transition ( $\Delta T=1$ ), so that the vector part, corresponding to the off-diagonal element of the isospin current, should vanish in the ideal limit of strict isospin invariance and  $q \rightarrow 0$ . In other words, we expect

$$\Gamma_V \approx [q^2 i\gamma_\mu - (m_\Sigma - m_\Lambda)q_\mu]F_{1V}(q^2) + \sigma_{\mu\nu} q_\nu F_{2V}(q^2). \quad (6.7)$$

In case the  $\Sigma-\Lambda$  parity is odd,<sup>25</sup> the vector and axial vector parts will interchange their roles. The vector part, which now looks like the axial vector current, would have the form

$$\begin{aligned} \Gamma_V \approx & [q^2 i\gamma_\mu\gamma_5 + (m_\Sigma + m_\Lambda)\gamma_5 q_\mu]F_{1V}(q^2) \\ & + i\gamma_5 \sigma_{\mu\nu} q_\nu F_{2V}(q^2). \end{aligned} \quad (6.8)$$

The axial vector part can similarly be put in the form

$$\begin{aligned} [i\gamma_\mu - q_\mu(m_\Sigma - m_\Lambda)/(q^2 + \mu_\pi^2)f(q^2)]F_{1A}(q^2) \\ + \sigma_{\mu\nu} q_\nu F_{2A}(q^2). \end{aligned} \quad (6.9)$$

But  $f(q^2)$  need not be  $\approx 1$  if the  $\Sigma-\Lambda$  mass difference is also due to the violation of the  $\gamma_5$  symmetry.

There are other processes for which the chirality conservation can be tested in a direct way. Although

<sup>23</sup> L. B. Okun', *Ann. Rev. Nuclear Sci.* **9**, 61 (1959); M. Gell-Mann, *Proceedings of the 1960 Annual International Conference on High-Energy Physics at Rochester* (Interscience Publishers, Inc., New York, 1960), p. 522.

<sup>24</sup> We have in this case three independent terms.

<sup>25</sup> See S. Barshay, *Phys. Rev. Letters* **1**, 97 (1958); Y. Nambu and J. J. Sakurai, *ibid.* **6**, 377 (1961).

extraordinary solutions are in general not eigenstates of chirality (even under strict  $\gamma_5$  invariance), the conservation law should still apply to the expectation values of chirality. In fact, we can express the chirality conservation law  $\langle X_i \rangle = \langle X_f \rangle$  for any reaction  $i \rightarrow f$ ; for example

$$\begin{aligned} p + \pi &\rightarrow p + \pi, \quad p + \pi + \pi', \text{ etc.,} \\ p + p &\rightarrow p + p, \quad p + p + \pi, \text{ etc.,} \end{aligned}$$

as a relation between the change of nucleon chirality and the magnitude of the pion production amplitude.

The ideas outlined in this section will be taken up in more detail elsewhere.

#### APPENDIX

We calculate here the renormalization of the axial vector (Gamow-Teller) coupling constant  $g_A$  for nuclear  $\beta$  decay under the following assumptions:

(1) Under strict  $\gamma_5$  invariance (Gürsey type), there is no renormalization, namely  $g_A = g_{A0}$  ( $= g_{V0} = g_V$ ), where  $g_{A0}$  is the bare coupling constant.

(2) The violation of the invariance gives rise to the finite pion mass as well as the deviation of the ratio  $R = g_A/g_{A0} = g_A/g_V$  from unity, so that there is a functional relation between the two quantities.

Let us first consider the isovector axial vector vertex  $\Gamma_A$  in the usual perturbation theory. In our model, it consists of various graphs, some of which are shown in Fig. 2(a) and (b). The "ladder" graphs 2(a) have been considered in I as well as in the present paper, since they are intimately related to the  $\gamma_5$  gauge transformation. In I (Appendix) we found that  $R > 1$  when both pseudoscalar and pseudovector type interactions are present.<sup>26</sup> The graphs 2(b) have not been considered yet. These will come into our consideration as soon as we take corresponding higher-order approximations for the self-energy, which was briefly discussed in I. The chain of bubbles in these graphs will act like a meson when there is such a dynamical pole [Fig. 2(c)].

The (divergent) renormalization effect due to intermediate mesons is always negative,<sup>27</sup> irrespective of the type of the meson, so that the effect of these meson-like bubble graphs is also expected to be similar. When the chain does not produce a pole, however, the effect can be opposite.

Combining all these effects, we have no way to predict the resultant magnitude and sign of the renormalization correction. So we simply assume these contributions to cancel out under strict  $\gamma_5$  invariance.

Next let us suppose that the invariance is slightly violated. This will cause changes in the propagators

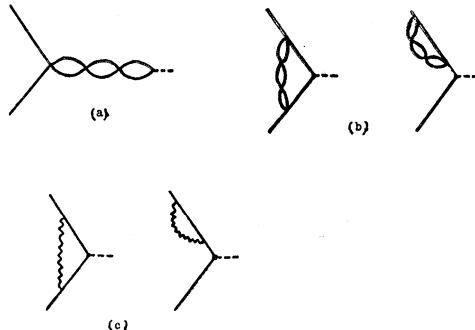


FIG. 2. Typical graphs considered in the evaluation of the axial vector vertex.

in all these graphs. Most of these changes are, however, quite small, being of the order of  $m^0/m \approx \mu^2/4m^2$ , as will be clear from the results of Sec. IV. The largest effect is naturally expected to come from the "pion" contribution in Fig. 2(b), as this is a change from zero mass (infinite range) to a finite one.

Let us accordingly take the effective pion graph from Fig. 2(c) with an arbitrary pion mass  $\mu$ . Call its contribution to the vertex renormalization (for zero momentum transfer)  $\Lambda(\mu)$ . Then according to the above assumption

$$\begin{aligned} R &= \Gamma_A(\mu_\pi)/g_A = \Gamma_A(\mu_\pi)/\Gamma_A(0) \\ &\approx 1 + \Lambda(\mu_\pi) - \Lambda(0). \end{aligned} \quad (\text{A1})$$

The difference  $\Lambda(\mu) - \Lambda(0)$  is convergent, which turns out to be

$$\begin{aligned} \Lambda(\mu) - \Lambda(0) &= \frac{G^2}{16\pi^2 m^2} \left\{ \left( 3 - \frac{5\mu^2}{2m^2} \right) \ln \frac{m^2}{\mu^2} - 5 \right. \\ &\quad \left. + \frac{16\mu}{\sqrt{3}m} \left[ \tan^{-1} \left( \frac{4m^2}{3\mu^2} - \frac{2}{3} \right) + \tan^{-1} \left( \frac{2}{3} \right) \right] \right\}, \end{aligned} \quad (\text{A2})$$

where  $G$  is the phenomenological pion coupling constant.

As was expected, this goes like  $(\mu^2/m^2) \ln(m^2/\mu^2)$  for small  $\mu$ , which is more important than the contributions from the neglected processes behaving like  $\mu^2/m^2$ .

With  $(G^2/4\pi)(\mu^2/4m^2) = f^2/4\pi = 0.08$ , Eq. (A2) gives

$$R - 1 = \begin{cases} 0.18 \\ 0.24. \end{cases} \quad (\text{A3})$$

The first figure is the entire contribution from Eq. (A2), while the second is the contribution from the leading logarithmic term alone. Experimentally,  $R$  is estimated to be  $\approx 1.25$ .<sup>28</sup>

<sup>26</sup> See also Z. Maki, Progr. Theoret. Phys. (Kyoto) **22**, 62 (1959).

<sup>27</sup> We owe Dr. J. de Swart the mathematical check on this point.

<sup>28</sup> M. T. Burgy, V. E. Krohn, T. B. Novey, G. R. Ringo, and V. L. Telegdi, Phys. Rev. **120**, 1829 (1961).



## A SCHEMATIC MODEL OF BARYONS AND MESONS \*

M. GELL-MANN

California Institute of Technology, Pasadena, California

Received 4 January 1964

If we assume that the strong interactions of baryons and mesons are correctly described in terms of the broken "eightfold way" 1-3), we are tempted to look for some fundamental explanation of the situation. A highly promised approach is the purely dynamical "bootstrap" model for all the strongly interacting particles within which one may try to derive isotopic spin and strangeness conservation and broken eightfold symmetry from self-consistency alone 4). Of course, with only strong interactions, the orientation of the asymmetry in the unitary space cannot be specified; one hopes that in some way the selection of specific components of the F-spin by electromagnetism and the weak interactions determines the choice of isotopic spin and hypercharge directions.

Even if we consider the scattering amplitudes of strongly interacting particles on the mass shell only and treat the matrix elements of the weak, electromagnetic, and gravitational interactions by means of dispersion theory, there are still meaningful and important questions regarding the algebraic properties of these interactions that have so far been discussed only by abstracting the properties from a formal field theory model based on fundamental entities 3) from which the baryons and mesons are built up.

If these entities were octets, we might expect the underlying symmetry group to be SU(8) instead of SU(3); it is therefore tempting to try to use unitary triplets as fundamental objects. A unitary triplet  $t$  consists of an isotopic singlet  $s$  of electric charge  $z$  (in units of  $e$ ) and an isotopic doublet  $(u, d)$  with charges  $z+1$  and  $z$  respectively. The anti-triplet  $\bar{t}$  has, of course, the opposite signs of the charges. Complete symmetry among the members of the triplet gives the exact eightfold way, while a mass difference, for example, between the isotopic doublet and singlet gives the first-order violation.

For any value of  $z$  and of triplet spin, we can construct baryon octets from a basic neutral baryon singlet  $b$  by taking combinations  $(btt\bar{t})$ ,  $(btt\bar{t}\bar{t})$ , etc. \*\*. From  $(btt\bar{t})$ , we get the representations 1 and 8, while from  $(btt\bar{t}\bar{t})$  we get 1, 8, 10, 10, and 27. In a similar way, meson singlets and octets can be made out of  $(t\bar{t})$ ,  $(tt\bar{t}\bar{t})$ , etc. The quantum num-

ber  $n_t - n_{\bar{t}}$  would be zero for all known baryons and mesons. The most interesting example of such a model is one in which the triplet has spin  $\frac{1}{2}$  and  $z = -1$ , so that the four particles  $d^-$ ,  $s^-$ ,  $u^0$  and  $b^0$  exhibit a parallel with the leptons.

A simpler and more elegant scheme can be constructed if we allow non-integral values for the charges. We can dispense entirely with the basic baryon  $b$  if we assign to the triplet  $t$  the following properties: spin  $\frac{1}{2}$ ,  $z = -\frac{1}{3}$ , and baryon number  $\frac{1}{3}$ . We then refer to the members  $u^{\frac{2}{3}}$ ,  $d^{-\frac{1}{3}}$ , and  $s^{-\frac{1}{3}}$  of the triplet as "quarks" 6)  $q$  and the members of the anti-triplet as anti-quarks  $\bar{q}$ . Baryons can now be constructed from quarks by using the combinations  $(qqq)$ ,  $(qqq\bar{q})$ , etc., while mesons are made out of  $(q\bar{q})$ ,  $(q\bar{q}\bar{q})$ , etc. It is assuming that the lowest baryon configuration  $(qqq)$  gives just the representations 1, 8, and 10 that have been observed, while the lowest meson configuration  $(q\bar{q})$  similarly gives just 1 and 8.

A formal mathematical model based on field theory can be built up for the quarks exactly as for  $p$ ,  $n$ ,  $\Lambda$  in the old Sakata model, for example 3) with all strong interactions ascribed to a neutral vector meson field interacting symmetrically with the three particles. Within such a framework, the electromagnetic current (in units of  $e$ ) is just

$$i\left\{\frac{2}{3}\bar{u}\gamma_\alpha u - \frac{1}{3}\bar{d}\gamma_\alpha d - \frac{1}{3}\bar{s}\gamma_\alpha s\right\}$$

or  $\mathcal{F}_{3\alpha} + \mathcal{F}_{8\alpha}/\sqrt{3}$  in the notation of ref. 3). For the weak current, we can take over from the Sakata model the form suggested by Gell-Mann and Lévy 7), namely  $i\bar{p}\gamma_\alpha(1+\gamma_5)(n\cos\theta + \Lambda\sin\theta)$ , which gives in the quark scheme the expression \*\*\*

$$i\bar{u}\gamma_\alpha(1+\gamma_5)(d\cos\theta + s\sin\theta)$$

\* Work supported in part by the U.S. Atomic Energy Commission.

\*\* This is similar to the treatment in ref. 1). See also ref. 5).

\*\*\* The parallel with  $i\bar{v}_e\gamma_\alpha(1+\gamma_5)e$  and  $i\bar{v}_\mu\gamma_\alpha(1+\gamma_5)\mu$  is obvious. Likewise, in the model with  $d^-$ ,  $s^-$ ,  $u^0$ , and  $b^0$  discussed above, we would take the weak current to be  $i(\bar{b}^0\cos\theta + \bar{u}^0\sin\theta)\gamma_\alpha(1+\gamma_5)s^- + i(\bar{u}^0\cos\theta - \bar{b}^0\sin\theta)\gamma_\alpha(1+\gamma_5)d^-$ . The part  $\Delta(n_t - n_{\bar{t}}) = 0$  is just  $i\bar{u}^0\gamma_\alpha(1+\gamma_5)(d^-\cos\theta + s^-\sin\theta)$ .

or, in the notation of ref. 3),

$$[\mathcal{F}_{1\alpha} + \mathcal{F}_{1\alpha}^5 + i(\mathcal{F}_{2\alpha} + \mathcal{F}_{2\alpha}^5)] \cos \theta \\ + [\mathcal{F}_{4\alpha} + \mathcal{F}_{4\alpha}^5 + i(\mathcal{F}_{5\alpha} + \mathcal{F}_{5\alpha}^5)] \sin \theta.$$

We thus obtain all the features of Cabibbo's picture<sup>8)</sup> of the weak current, namely the rules  $|\Delta I| = 1$ ,  $\Delta Y = 0$  and  $|\Delta I| = \frac{1}{2}$ ,  $\Delta Y/\Delta Q = +1$ , the conserved  $\Delta Y = 0$  current with coefficient  $\cos \theta$ , the vector current in general as a component of the current of the F-spin, and the axial vector current transforming under SU(3) as the same component of another octet. Furthermore, we have<sup>3)</sup> the equal-time commutation rules for the fourth components of the currents:

$$[\mathcal{F}_{j4}(x) \pm \mathcal{F}_{j4}^5(x), \mathcal{F}_{k4}(x') \pm \mathcal{F}_{k4}^5(x')] = \\ - 2 f_{jkl} [\mathcal{F}_{l4}(x) \pm \mathcal{F}_{l4}^5(x)] \delta(x-x'), \\ [\mathcal{F}_{j4}(x) \pm \mathcal{F}_{j4}^5(x), \mathcal{F}_{k4}(x') \mp \mathcal{F}_{k4}^5(x')] = 0,$$

$i = 1, \dots, 8$ , yielding the group  $SU(3) \times SU(3)$ . We can also look at the behaviour of the energy density  $\theta_{44}(x)$  (in the gravitational interaction) under equal-time commutation with the operators  $\mathcal{F}_{j4}(x') \pm \mathcal{F}_{j4}^5(x')$ . That part which is non-invariant under the group will transform like particular representations of  $SU(3) \times SU(3)$ , for example like  $(3, \bar{3})$  and  $(\bar{3}, 3)$  if it comes just from the masses of the quarks.

All these relations can now be abstracted from the field theory model and used in a dispersion theory treatment. The scattering amplitudes for strongly interacting particles on the mass shell are assumed known; there is then a system of linear dispersion relations for the matrix elements of the weak currents (and also the electromagnetic and gravitational interactions) to lowest order in these interactions. These dispersion relations, unsubtracted and supplemented by the non-linear commutation rules abstracted from the field theory, may be powerful enough to determine all the matrix elements of the weak currents, including the effective strengths of the axial vector current matrix elements compared with those of the vector current.

It is fun to speculate about the way quarks would behave if they were physical particles of finite mass

(instead of purely mathematical entities as they would be in the limit of infinite mass). Since charge and baryon number are exactly conserved, one of the quarks (presumably  $u^{\frac{2}{3}}$  or  $d^{-\frac{1}{3}}$ ) would be absolutely stable\*, while the other member of the doublet would go into the first member very slowly by  $\beta$ -decay or K-capture. The isotopic singlet quark would presumably decay into the doublet by weak interactions, much as  $\Lambda$  goes into  $N$ . Ordinary matter near the earth's surface would be contaminated by stable quarks as a result of high energy cosmic ray events throughout the earth's history, but the contamination is estimated to be so small that it would never have been detected. A search for stable quarks of charge  $-\frac{1}{3}$  or  $+\frac{2}{3}$  and/or stable di-quarks of charge  $-\frac{2}{3}$  or  $+\frac{1}{3}$  or  $+\frac{4}{3}$  at the highest energy accelerators would help to reassure us of the non-existence of real quarks.

These ideas were developed during a visit to Columbia University in March 1963; the author would like to thank Professor Robert Serber for stimulating them.

#### References

- 1) M. Gell-Mann, California Institute of Technology Synchrotron Laboratory Report CTSL-20 (1961).
- 2) Y. Ne'eman, Nuclear Phys. 26 (1961) 222.
- 3) M. Gell-Mann, Phys. Rev. 125 (1962) 1067.
- 4) E.g.: R. H. Capps, Phys. Rev. Letters 10 (1963) 312; R. E. Cutkosky, J. Kalckar and P. Tarjanne, Physics Letters 1 (1962) 93; E. Abers, F. Zachariasen and A. C. Zemach, Phys. Rev. 132 (1963) 1831; S. Glashow, Phys. Rev. 130 (1963) 2132; R. E. Cutkosky and P. Tarjanne, Phys. Rev. 132 (1963) 1354.
- 5) P. Tarjanne and V. L. Teplitz, Phys. Rev. Letters 11 (1963) 447.
- 6) James Joyce, *Finnegan's Wake* (Viking Press, New York, 1939) p. 383.
- 7) M. Gell-Mann and M. Lévy, Nuovo Cimento 16 (1960) 705.
- 8) N. Cabibbo, Phys. Rev. Letters 10 (1963) 531.

\* There is the alternative possibility that the quarks are unstable under decay into baryon plus anti-di-quark or anti-baryon plus quadri-quark. In any case, some particle of fractional charge would have to be absolutely stable.

\* \* \* \*

## BROKEN SYMMETRIES AND THE MASSES OF GAUGE BOSONS

Peter W. Higgs

Tait Institute of Mathematical Physics, University of Edinburgh, Edinburgh, Scotland  
(Received 31 August 1964)

In a recent note<sup>1</sup> it was shown that the Goldstone theorem,<sup>2</sup> that Lorentz-covariant field theories in which spontaneous breakdown of symmetry under an internal Lie group occurs contain zero-mass particles, fails if and only if the conserved currents associated with the internal group are coupled to gauge fields. The purpose of the present note is to report that, as a consequence of this coupling, the spin-one quanta of some of the gauge fields acquire mass; the longitudinal degrees of freedom of these particles (which would be absent if their mass were zero) go over into the Goldstone bosons when the coupling tends to zero. This phenomenon is just the relativistic analog of the plasmon phenomenon to which Anderson<sup>3</sup> has drawn attention: that the scalar zero-mass excitations of a superconducting neutral Fermi gas become longitudinal plasmon modes of finite mass when the gas is charged.

The simplest theory which exhibits this behavior is a gauge-invariant version of a model used by Goldstone<sup>2</sup> himself: Two real<sup>4</sup> scalar fields  $\varphi_1, \varphi_2$  and a real vector field  $A_\mu$  interact through the Lagrangian density

$$L = -\frac{1}{2}(\nabla\varphi_1)^2 - \frac{1}{2}(\nabla\varphi_2)^2 - V(\varphi_1^2 + \varphi_2^2) - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}, \quad (1)$$

where

$$\nabla_\mu\varphi_1 = \partial_\mu\varphi_1 - eA_\mu\varphi_2,$$

$$\nabla_\mu\varphi_2 = \partial_\mu\varphi_2 + eA_\mu\varphi_1,$$

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu,$$

$e$  is a dimensionless coupling constant, and the metric is taken as  $-+++$ .  $L$  is invariant under simultaneous gauge transformations of the first kind on  $\varphi_1 \pm i\varphi_2$  and of the second kind on  $A_\mu$ . Let us suppose that  $V'(\varphi_0^2) = 0, V''(\varphi_0^2) > 0$ ; then spontaneous breakdown of U(1) symmetry occurs. Consider the equations [derived from (1) by treating  $\Delta\varphi_1, \Delta\varphi_2$ , and  $A_\mu$  as small quantities] governing the propagation of small oscillations

about the "vacuum" solution  $\varphi_1(x) = 0, \varphi_2(x) = \varphi_0$ :

$$\partial^\mu\{\partial_\mu(\Delta\varphi_1) - e\varphi_0 A_\mu\} = 0, \quad (2a)$$

$$\{\partial^2 - 4\varphi_0^2 V''(\varphi_0^2)\}(\Delta\varphi_2) = 0, \quad (2b)$$

$$\partial_\nu F^{\mu\nu} = e\varphi_0\{\partial^\mu(\Delta\varphi_1) - e\varphi_0 A_\mu\}. \quad (2c)$$

Equation (2b) describes waves whose quanta have (bare) mass  $2\varphi_0\{V''(\varphi_0^2)\}^{1/2}$ ; Eqs. (2a) and (2c) may be transformed, by the introduction of new variables

$$\begin{aligned} B_\mu &= A_\mu - (e\varphi_0)^{-1}\partial_\mu(\Delta\varphi_1), \\ G_{\mu\nu} &= \partial_\mu B_\nu - \partial_\nu B_\mu = F_{\mu\nu}, \end{aligned} \quad (3)$$

into the form

$$\partial_\mu B^\mu = 0, \quad \partial_\nu G^{\mu\nu} + e^2\varphi_0^2 B^\mu = 0. \quad (4)$$

Equation (4) describes vector waves whose quanta have (bare) mass  $e\varphi_0$ . In the absence of the gauge field coupling ( $e = 0$ ) the situation is quite different: Equations (2a) and (2c) describe zero-mass scalar and vector bosons, respectively. In passing, we note that the right-hand side of (2c) is just the linear approximation to the conserved current: It is linear in the vector potential, gauge invariance being maintained by the presence of the gradient term.<sup>5</sup>

When one considers theoretical models in which spontaneous breakdown of symmetry under a semisimple group occurs, one encounters a variety of possible situations corresponding to the various distinct irreducible representations to which the scalar fields may belong; the gauge field always belongs to the adjoint representation.<sup>6</sup> The model of the most immediate interest is that in which the scalar fields form an octet under SU(3): Here one finds the possibility of two nonvanishing vacuum expectation values, which may be chosen to be the two  $Y = 0, I_3 = 0$  members of the octet.<sup>7</sup> There are two massive scalar bosons with just these quantum numbers; the remaining six components of the scalar octet combine with the corresponding components of the gauge-field octet to describe

massive vector bosons. There are two  $I=\frac{1}{2}$  vector doublets, degenerate in mass between  $Y=\pm 1$  but with an electromagnetic mass splitting between  $I_3=\pm \frac{1}{2}$ , and the  $I_3=\pm 1$  components of a  $Y=0$ ,  $I=1$  triplet whose mass is entirely electromagnetic. The two  $Y=0$ ,  $I=0$  gauge fields remain massless: This is associated with the residual unbroken symmetry under the Abelian group generated by  $Y$  and  $I_3$ . It may be expected that when a further mechanism (presumably related to the weak interactions) is introduced in order to break  $Y$  conservation, one of these gauge fields will acquire mass, leaving the photon as the only massless vector particle. A detailed discussion of these questions will be presented elsewhere.

It is worth noting that an essential feature of the type of theory which has been described in this note is the prediction of incomplete multiplets of scalar and vector bosons.<sup>8</sup> It is to be expected that this feature will appear also in theories in which the symmetry-breaking scalar fields are not elementary dynamic variables but bilinear combinations of Fermi fields.<sup>9</sup>

<sup>1</sup>P. W. Higgs, to be published.

<sup>2</sup>J. Goldstone, Nuovo Cimento 19, 154 (1961);  
J. Goldstone, A. Salam, and S. Weinberg, Phys. Rev. 127, 965 (1962).

<sup>3</sup>P. W. Anderson, Phys. Rev. 130, 439 (1963).

<sup>4</sup>In the present note the model is discussed mainly in classical terms; nothing is proved about the quantized theory. It should be understood, therefore, that the conclusions which are presented concerning the masses of particles are conjectures based on the quantization of linearized classical field equations. However, essentially the same conclusions have been reached independently by F. Englert and R. Brout, Phys. Rev. Letters 13, 321 (1964): These authors discuss the same model quantum mechanically in lowest order perturbation theory about the self-consistent vacuum.

<sup>5</sup>In the theory of superconductivity such a term arises from collective excitations of the Fermi gas.

<sup>6</sup>See, for example, S. L. Glashow and M. Gell-Mann, Ann. Phys. (N.Y.) 15, 437 (1961).

<sup>7</sup>These are just the parameters which, if the scalar octet interacts with baryons and mesons, lead to the Gell-Mann-Okubo and electromagnetic mass splittings: See S. Coleman and S. L. Glashow, Phys. Rev. 134, B671 (1964).

<sup>8</sup>Tentative proposals that incomplete SU(3) octets of scalar particles exist have been made by a number of people. Such a rôle, as an isolated  $Y=\pm 1$ ,  $I=\frac{1}{2}$  state, was proposed for the  $\kappa$  meson (725 MeV) by Y. Nambu and J. J. Sakurai, Phys. Rev. Letters 11, 42 (1963). More recently the possibility that the  $\sigma$  meson (385 MeV) may be the  $Y=I=0$  member of an incomplete octet has been considered by L. M. Brown, Phys. Rev. Letters 13, 42 (1964).

<sup>9</sup>In the theory of superconductivity the scalar fields are associated with fermion pairs; the doubly charged excitation responsible for the quantization of magnetic flux is then the surviving member of a U(1) doublet.

## Inhomogeneous Electron Gas\*

P. HOHENBERG†

École Normale Supérieure, Paris, France

AND

W. KOHN‡

École Normale Supérieure, Paris, France and Faculté des Sciences, Orsay, France

and

University of California at San Diego, La Jolla, California

(Received 18 June 1964)

This paper deals with the ground state of an interacting electron gas in an external potential  $v(\mathbf{r})$ . It is proved that there exists a universal functional of the density,  $F[n(\mathbf{r})]$ , independent of  $v(\mathbf{r})$ , such that the expression  $E \equiv \int v(\mathbf{r})n(\mathbf{r})d\mathbf{r} + F[n(\mathbf{r})]$  has as its minimum value the correct ground-state energy associated with  $v(\mathbf{r})$ . The functional  $F[n(\mathbf{r})]$  is then discussed for two situations: (1)  $n(\mathbf{r}) = n_0 + \tilde{n}(\mathbf{r})$ ,  $\tilde{n}/n_0 < < 1$ , and (2)  $n(\mathbf{r}) = \varphi(\mathbf{r}/r_0)$  with  $\varphi$  arbitrary and  $r_0 \rightarrow \infty$ . In both cases  $F$  can be expressed entirely in terms of the correlation energy and linear and higher order electronic polarizabilities of a uniform electron gas. This approach also sheds some light on generalized Thomas-Fermi methods and their limitations. Some new extensions of these methods are presented.

## INTRODUCTION

DURING the last decade there has been considerable progress in understanding the properties of a homogeneous interacting electron gas.<sup>1</sup> The point of view has been, in general, to regard the electrons as similar to a collection of noninteracting particles with the important additional concept of collective excitations.

On the other hand, there has been in existence since the 1920's a different approach, represented by the Thomas-Fermi method<sup>2</sup> and its refinements, in which the electronic density  $n(\mathbf{r})$  plays a central role and in which the system of electrons is pictured more like a classical liquid. This approach has been useful, up to now, for simple though crude descriptions of inhomogeneous systems like atoms and impurities in metals.

Lately there have been also some important advances along this second line of approach, such as the work of Kompaneets and Pavlovskii,<sup>3</sup> Kirzhnits,<sup>4</sup> Lewis,<sup>5</sup> Baraff and Borowitz,<sup>6</sup> Baraff,<sup>7</sup> and DuBois and Kivelson.<sup>8</sup> The present paper represents a contribution in the same area.

In Part I, we develop an exact formal variational principle for the ground-state energy, in which the density  $n(\mathbf{r})$  is the variable function. Into this principle enters a universal functional  $F[n(\mathbf{r})]$ , which applies to all electronic systems in their ground state no matter what the external potential is. The main objective of

theoretical considerations is a description of this functional. Once known, it is relatively easy to determine the ground-state energy in a given external potential.

In Part II, we obtain an expression for  $F[n]$  when  $n$  deviates only slightly from uniformity, i.e.,  $n(\mathbf{r}) = n_0 + \tilde{n}(\mathbf{r})$ , with  $\tilde{n}/n_0 \rightarrow 0$ . In this case  $F[n]$  is entirely expressible in terms of the exact ground-state energy and the exact electronic polarizability  $\alpha(g)$  of a uniform electron gas. This procedure will describe correctly the long-range Friedel charge oscillations<sup>9</sup> set up by a localized perturbation. All previous refinements of the Thomas-Fermi method have failed to include these.

In Part III we consider the case of a slowly varying, but *not necessarily almost constant* density,  $n(\mathbf{r}) = \varphi(\mathbf{r}/r_0)$ ,  $r_0 \rightarrow \infty$ . For this case we derive an expansion of  $F[n]$  in successive orders of  $r_0^{-1}$  or, equivalently of the gradient operator  $\nabla$  acting on  $n(\mathbf{r})$ . The expansion coefficients are again expressible in terms of the exact ground-state energy and the exact linear, quadratic, etc., electric response functions of a uniform electron gas to an external potential  $v(\mathbf{r})$ . In this way we recover, quite simply, all previously developed refinements of the Thomas-Fermi method and are able to carry them somewhat further. Comparison of this case with the nearly uniform one, discussed in Part II, also reveals why the gradient expansion is intrinsically incapable of properly describing the Friedel oscillations or the radial oscillations of the electronic density in an atom which reflect the electronic shell structure. A partial summation of the gradient expansion can be carried out (Sec. III.4), but its usefulness has not yet been tested.

## I. EXACT GENERAL FORMULATION

## 1. The Density as Basic Variable

We shall be considering a collection of an arbitrary number of electrons, enclosed in a large box and moving

\* Supported in part by the U. S. Office of Naval Research.

† NATO Post Doctoral Fellow.

‡ Guggenheim Fellow.

<sup>1</sup> For a review see, for example, D. Pines, *Elementary Excitations in Solids* (W. A. Benjamin Inc., New York, 1963).

<sup>2</sup> For a review of work up to 1956, see N. H. March, *Advan. Phys.* **6**, 1 (1957).

<sup>3</sup> A. S. Kompaneets and E. S. Pavlovskii, *Zh. Eksperim. i Teor. Fiz.* **31**, 427 (1956) [English transl.: Soviet Phys.—JETP **4**, 328 (1957)].

<sup>4</sup> D. A. Kirzhnits, *Zh. Eksperim. i. Teor. Fiz.* **32**, 115 (1957) [English transl.: Soviet Phys.—JETP **5**, 64 (1957)].

<sup>5</sup> H. W. Lewis, *Phys. Rev.* **111**, 1554 (1958).

<sup>6</sup> G. A. Baraff and S. Borowitz, *Phys. Rev.* **121**, 1704 (1961).

<sup>7</sup> G. A. Baraff, *Phys. Rev.* **123**, 2087 (1961).

<sup>8</sup> D. F. Du Bois and M. G. Kivelson, *Phys. Rev.* **127**, 1182 (1962).

<sup>9</sup> J. Friedel, *Phil. Mag.* **43**, 153 (1952).

under the influence of an external potential  $v(\mathbf{r})$  and the mutual Coulomb repulsion. The Hamiltonian has the form

$$H = T + V + U, \quad (1)$$

where<sup>10</sup>

$$T = \frac{1}{2} \int \nabla \psi^*(\mathbf{r}) \nabla \psi(\mathbf{r}) d\mathbf{r}, \quad (2)$$

$$V = \int v(\mathbf{r}) \psi^*(\mathbf{r}) \psi(\mathbf{r}) d\mathbf{r}, \quad (3)$$

$$U = \frac{1}{2} \int \frac{1}{|\mathbf{r} - \mathbf{r}'|} \psi^*(\mathbf{r}) \psi^*(\mathbf{r}') \psi(\mathbf{r}') \psi(\mathbf{r}) d\mathbf{r} d\mathbf{r}'. \quad (4)$$

We shall in all that follows assume for simplicity that we are only dealing with situations in which the ground state is nondegenerate. We denote the electronic density in the ground state  $\Psi$  by

$$n(\mathbf{r}) \equiv (\Psi, \psi^*(\mathbf{r}) \psi(\mathbf{r}) \Psi), \quad (5)$$

which is clearly a functional of  $v(\mathbf{r})$ .

We shall now show that conversely  $v(\mathbf{r})$  is a unique functional of  $n(\mathbf{r})$ , apart from a trivial additive constant.

The proof proceeds by *reductio ad absurdum*. Assume that another potential  $v'(\mathbf{r})$ , with ground state  $\Psi'$  gives rise to the same density  $n(\mathbf{r})$ . Now clearly [unless  $v'(\mathbf{r}) - v(\mathbf{r}) = \text{const}$ ]  $\Psi'$  cannot be equal to  $\Psi$  since they satisfy different Schrödinger equations. Hence, if we denote the Hamiltonian and ground-state energies associated with  $\Psi$  and  $\Psi'$  by  $H$ ,  $H'$  and  $E$ ,  $E'$ , we have by the minimal property of the ground state,

$$E' = (\Psi', H' \Psi') < (\Psi, H' \Psi) = (\Psi, (H + V' - V) \Psi),$$

so that

$$E' < E + \int [v'(\mathbf{r}) - v(\mathbf{r})] n(\mathbf{r}) d\mathbf{r}. \quad (6)$$

Interchanging primed and unprimed quantities, we find in exactly the same way that

$$E < E' + \int [v(\mathbf{r}) - v'(\mathbf{r})] n(\mathbf{r}) d\mathbf{r}. \quad (7)$$

Addition of (6) and (7) leads to the inconsistency

$$E + E' < E + E'. \quad (8)$$

Thus  $v(\mathbf{r})$  is (to within a constant) a unique functional of  $n(\mathbf{r})$ ; since, in turn,  $v(\mathbf{r})$  fixes  $H$  we see that the full many-particle ground state is a unique functional of  $n(\mathbf{r})$ .

## 2. The Variational Principle

Since  $\Psi$  is a functional of  $n(\mathbf{r})$ , so is evidently the kinetic and interaction energy. We therefore define

$$F[n(\mathbf{r})] \equiv (\Psi, (T + U) \Psi), \quad (9)$$

<sup>10</sup> Atomic units are used.

where  $F[n]$  is a universal functional, valid for any number of particles<sup>11</sup> and *any* external potential. This functional plays a central role in the present paper.

With its aid we define, for a given potential  $v(\mathbf{r})$ , the energy functional

$$E_v[n] \equiv \int v(\mathbf{r}) n(\mathbf{r}) d\mathbf{r} + F[n]. \quad (10)$$

Clearly, for the correct  $n(\mathbf{r})$ ,  $E_v[n]$  equals the ground-state energy  $E$ .

We shall now show that  $E_v[n]$  assumes its minimum value for the correct  $n(\mathbf{r})$ , if the admissible functions are restricted by the condition

$$N[n] \equiv \int n(\mathbf{r}) d\mathbf{r} = N. \quad (11)$$

It is well known that for a system of  $N$  particles, the energy functional of  $\Psi'$

$$\mathcal{E}_v[\Psi'] \equiv (\Psi', V\Psi') + (\Psi', (T + U)\Psi') \quad (12)$$

has a minimum at the correct ground state  $\Psi$ , relative to arbitrary variations of  $\Psi'$  in which the number of particles is kept constant. In particular, let  $\Psi'$  be the ground state associated with a different external potential  $v'(\mathbf{r})$ . Then, by (12) and (9)

$$\begin{aligned} \mathcal{E}_v[\Psi'] &= \int v(\mathbf{r}) n'(\mathbf{r}) d\mathbf{r} + F[n'], \\ &> \mathcal{E}_v[\Psi] = \int v(\mathbf{r}) n(\mathbf{r}) d\mathbf{r} + F[n]. \end{aligned} \quad (13)$$

Thus the minimal property of (10) is established relative to all density functions  $n'(\mathbf{r})$  associated with some other external potential  $v'(\mathbf{r})$ .<sup>12</sup>

If  $F[n]$  were a known and sufficiently simple functional of  $n$ , the problem of determining the ground-state energy and density in a given external potential would be rather easy since it requires merely the minimization of a functional of the three-dimensional density function. The major part of the complexities of the many-electron problems are associated with the determination of the universal functional  $F[n]$ .

## 3. Transformation of the Functional $F[n]$

Because of the long range of the Coulomb interaction, it is for most purposes convenient to separate out from

<sup>11</sup> This is obvious since the number of particles is itself a simple functional of  $n(\mathbf{r})$ .

<sup>12</sup> We cannot prove whether an arbitrary positive density distribution  $n'(\mathbf{r})$ , which satisfies the condition  $\int n'(\mathbf{r}) d\mathbf{r} = \text{integer}$ , can be realized by *some* external potential  $v'(\mathbf{r})$ . Clearly, to first order in  $\tilde{n}(\mathbf{r})$ , any distribution of the form  $n'(\mathbf{r}) = n_0 + \tilde{n}(\mathbf{r})$  can be so realized and we believe that in fact all, except some pathological distributions, can be realized.

$F[n]$  the classical Coulomb energy and write

$$F[n] = \frac{1}{2} \int \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}d\mathbf{r}' + G[n], \quad (14)$$

so that  $E_v[n]$  becomes

$$E_v[n] = \int v(\mathbf{r})n(\mathbf{r})d\mathbf{r} + \frac{1}{2} \int \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}d\mathbf{r}' + G[n], \quad (15)$$

where  $G[n]$  is a universal functional like  $F[n]$ .

Now from the definition of  $F[n]$ , Eq. (9), and  $G[n]$ , Eq. (14), we see that

$$G[n] = \frac{1}{2} \int \nabla_{\mathbf{r}} \nabla_{\mathbf{r}'} n_1(\mathbf{r}, \mathbf{r}') |_{\mathbf{r}=\mathbf{r}'} d\mathbf{r} + \frac{1}{2} \int \frac{C_2(\mathbf{r}, \mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}d\mathbf{r}'. \quad (16)$$

Here  $n_1(\mathbf{r}, \mathbf{r}')$  is the one-particle density matrix; and  $C_2(\mathbf{r}, \mathbf{r}')$  is the two-particle correlation function defined in terms of the one- and two-particle density matrices as

$$C_2(\mathbf{r}, \mathbf{r}') = n_2(\mathbf{r}, \mathbf{r}'; \mathbf{r}, \mathbf{r}') - n_1(\mathbf{r}, \mathbf{r})n_1(\mathbf{r}', \mathbf{r}'). \quad (17)$$

Of course  $n_1(\mathbf{r}, \mathbf{r}) \equiv n(\mathbf{r})$ .

From (16) we see that we can define an energy-density functional

$$g_r[n] = \frac{1}{2} \nabla_{\mathbf{r}} \nabla_{\mathbf{r}'} n_1(\mathbf{r}, \mathbf{r}') |_{\mathbf{r}=\mathbf{r}'} + \frac{1}{2} \int \frac{C_2(\mathbf{r}-\mathbf{r}'/2; \mathbf{r}+\mathbf{r}'/2)}{|\mathbf{r}'|} d\mathbf{r}' \quad (18)$$

such that

$$G[n] = \int g_r[n] d\mathbf{r}. \quad (19)$$

The fact that  $g_r[n]$  is a functional of  $n$  follows of course from the fact that  $\Psi$  and hence  $n_1$  and  $n_2$  are.

It should be remarked, that while  $G[n]$  is a unique functional of  $n$ ,  $g_r[n]$  is of course not the only possible energy-density functional. Clearly the functionals

$$\tilde{g}_r[n] = g_r[n] + \sum_{i=1}^3 \frac{\partial}{\partial x_i} h_r^{(i)}[n], \quad (20)$$

where the  $h_r^{(i)}$  are entirely arbitrary, give equivalent results when used in conjunction with (19).

The following sections deal with  $G[n]$  and  $g_r[n]$  in some simple cases.

## II. THE GAS OF ALMOST CONSTANT DENSITY

### 1. Form of the Functionals $G[n]$ and $\bar{g}_r[n]$

We consider here a gas whose density has the form

$$n(\mathbf{r}) = n_0 + \tilde{n}(\mathbf{r}), \quad (21)$$

with

$$\tilde{n}(\mathbf{r})/n_0 \ll 1 \quad (22)$$

and

$$\int \tilde{n}(\mathbf{r}) d\mathbf{r} = 0. \quad (23)$$

Here we clearly must have a formal expansion of the following sort:

$$G[n] = G[n_0] + \int K(\mathbf{r}-\mathbf{r}') \tilde{n}(\mathbf{r}) \tilde{n}(\mathbf{r}') d\mathbf{r}d\mathbf{r}' + \int L(\mathbf{r}, \mathbf{r}', \mathbf{r}'') \tilde{n}(\mathbf{r}) \tilde{n}(\mathbf{r}') \tilde{n}(\mathbf{r}'') d\mathbf{r}d\mathbf{r}'d\mathbf{r}'' + \dots \quad (24)$$

In this equation there is no term linear in  $\tilde{n}(\mathbf{r})$  since by translational invariance the coefficient of  $\tilde{n}(\mathbf{r})$  would be independent of  $\mathbf{r}$  leading to zero, by (23). The kernel appearing in the quadratic term is a functional of  $|\mathbf{r}-\mathbf{r}'|$  only and may therefore be written as

$$K(\mathbf{r}-\mathbf{r}') = (1/\Omega) \sum_{\mathbf{q}} K(\mathbf{q}) e^{-i\mathbf{q} \cdot (\mathbf{r}-\mathbf{r}')}. \quad (25)$$

The higher order terms will not be further discussed here.

One may also quite trivially introduce a density function

$$\bar{g}_r[n] = g_0(n_0) + \int K(\mathbf{r}') \tilde{n}(\mathbf{r} + \frac{1}{2}\mathbf{r}') \tilde{n}(\mathbf{r} - \frac{1}{2}\mathbf{r}') d\mathbf{r}' + \dots, \quad (26)$$

where  $g_0(n_0)$  is the density function of a uniform gas of electron density  $n_0$  (kinetic, exchange, and correlation energy).

### 2. Expression of the Kernel $K$ in Terms of the Electronic Polarizability

We shall now see that the kernel  $K$  appearing in Eqs. (24) and (26) is completely and exactly expressible in terms of the electronic polarizability  $\alpha(q)$ . The latter is defined as follows: Consider an electron gas of mean density  $n_0$  in a background of uniform charge plus a small additional positive external-charge density

$$n_{\text{ext}}(\mathbf{r}) = (\lambda/\Omega) \sum_{\mathbf{q}} a(\mathbf{q}) e^{-i\mathbf{q} \cdot \mathbf{r}}. \quad (27)$$

Write the electronic density, to first order in  $\lambda$ , as

$$n(\mathbf{r}) = n_0 + (\lambda/\Omega) \sum_{\mathbf{q}} b_1(\mathbf{q}) e^{-i\mathbf{q} \cdot \mathbf{r}}. \quad (28)$$

Then

$$\alpha(q) = b_1(\mathbf{q})/a(\mathbf{q}). \quad (29)$$

Let us now define the operator

$$\rho_{\mathbf{q}} \equiv \sum_{\mathbf{k}} c_{\mathbf{k}-\mathbf{q}}^* c_{\mathbf{k}}, \quad (30)$$

where  $c_{\mathbf{k}}^*$ ,  $c_{\mathbf{k}}$  are the usual creation and annihilation operators. Then, by first-order perturbation theory,

$$b_1(\mathbf{q}) = -(8\pi)^{-1} \frac{a(q)}{q^2} \sum_n \frac{(0| \rho_{\mathbf{q}} | n)(n| \rho_{-\mathbf{q}} | 0)}{E_0 - E_n}, \quad (31)$$

so that

$$\alpha(q) = \frac{-8\pi}{q^2} \sum_n \frac{(0|\rho_{\mathbf{q}}|n)(n|\rho_{-\mathbf{q}}|0)}{E_0 - E_n}. \quad (32)$$

Next we express the change of energy in terms of  $\alpha(q)$ . By second-order perturbation theory we have

$$\begin{aligned} E &= E_0 + \frac{\lambda^2(4\pi)^2}{\Omega} \sum_{\mathbf{q}} \frac{|\alpha(\mathbf{q})|^2}{q^4} \sum_n \frac{(0|\rho_{\mathbf{q}}|n)(n|\rho_{-\mathbf{q}}|0)}{E_0 - E_n}, \\ &= E_0 - \frac{\lambda^2 2\pi}{\Omega} \sum_{\mathbf{q}} \frac{|\alpha(\mathbf{q})|^2}{q^2} \alpha(q), \\ &= E_0 - \frac{\lambda^2 2\pi}{\Omega} \sum_{\mathbf{q}} \frac{|b_1(\mathbf{q})|^2}{\alpha(q) q^2}. \end{aligned} \quad (33)$$

On the other hand, combining Eqs. (15), (24), (25), and (28) gives

$$\begin{aligned} E &= \int v(\mathbf{r}) n(\mathbf{r}) + \frac{1}{2} \int \frac{n(\mathbf{r}) n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r} d\mathbf{r}' + G[n] \\ &= E_0 - \frac{\lambda^2 4\pi}{\Omega} \sum_{\mathbf{q}} \frac{|b_1(\mathbf{q})|^2}{\alpha(q) q^2} + \frac{\lambda^2 2\pi}{\Omega} \sum_{\mathbf{q}} \frac{|b_1(\mathbf{q})|^2}{q^2} \\ &\quad + \frac{\lambda^3}{\Omega} \sum_{\mathbf{q}} K(\mathbf{q}) |b_1(\mathbf{q})|^2. \end{aligned} \quad (34)$$

Comparison of Eqs. (33) and (34) gives

$$K(q) = \frac{2\pi}{q^2} \left[ \frac{1}{\alpha(q)} - 1 \right]. \quad (35)$$

Equivalently, in terms of the dielectric constant,

$$\epsilon(q) = \frac{1}{1 - \alpha(q)}, \quad (36)$$

we may write

$$K(q) = \frac{2\pi}{q^2} \frac{1}{\epsilon(q) - 1}. \quad (37)$$

### 3. The Nature of the Kernel $K$

The polarizability  $\alpha(q)$  has the following properties, as function of  $q$  (see Fig. 1)

$$q \rightarrow 0: \quad \alpha(q) = 1 + c_2 q^2 + c_4 q^4 + \dots; \quad (38)$$

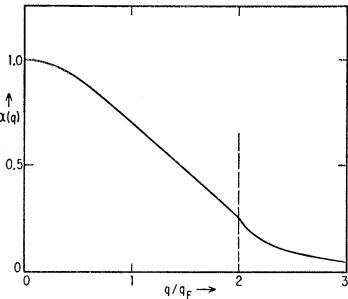
$$q \rightarrow 2k_F: \quad d\alpha/dq \rightarrow -\infty; \quad (39)$$

$$q \rightarrow \infty: \quad \alpha(q) \rightarrow \text{const}/q^4. \quad (40)$$

These general properties are exemplified by the random-phase approximation in which

$$\alpha(q) = [1 + (q^2/k_F^2) S(q)]^{-1} \quad (41)$$

FIG. 1. Behavior of the electronic polarizability  $\alpha(q)$ , as function of  $q$  (electronic density =  $4 \times 10^{23} \text{ cm}^{-3}$ ).



where  $k_F$  is the Thomas-Fermi screening constant,

$$k_F \equiv (4k_F)^{1/2} \quad (42)$$

and

$$S(q) \equiv \left[ \frac{1}{2} + \frac{k_F}{2q} \left( 1 - \frac{q^2}{4k_F^2} \right) \ln \left| \frac{q+2k_F}{q-2k_F} \right| \right]^{-1}. \quad (43)$$

This gives for  $K(q)$ , by (35),

$$q \rightarrow 0: \quad K(q) = 2\pi[-c_2 + (c_2^2 - c_4)q^2 + \dots]; \quad (44)$$

$$q \rightarrow 2k_F: \quad dK/dq \rightarrow +\infty; \quad (45)$$

$$q \rightarrow \infty: \quad K(q) \rightarrow \text{const} \propto q^2. \quad (46)$$

(See Fig. 2.)

The power-series expansion of  $K(q)$ , (43), leads to

$$K(\mathbf{r}) = 2\pi[-c_2 + (c_2^2 - c_4)\nabla^2 + \dots]\delta(\mathbf{r}), \quad (47)$$

which in turn gives

$$\begin{aligned} G[n] &= G[n_0] + 2\pi \left[ -c_2 \int \tilde{n}(\mathbf{r})^2 d\mathbf{r} \right. \\ &\quad \left. + (c_2^2 - c_4) \int |\nabla \tilde{n}(\mathbf{r})|^2 d\mathbf{r} + \dots \right], \end{aligned} \quad (48)$$

i.e., a gradient expansion.

At this point an important remark must be made. One of the most significant features of  $K(q)$  is its singularity at  $q = 2k_F$ . This is responsible for the long-range Friedel oscillations<sup>13</sup> in  $K(\mathbf{r})$ ,

$$r \rightarrow \infty: \quad K(r) \sim \text{const} \cos(2k_F r + \delta)/r^3. \quad (49)$$

These obviously lie outside the framework of the power-series expansion (44) of  $K(q)$  and hence outside the gradient expansion (49) of  $G[n]$ . This explains why neither the original Thomas-Fermi method [which for the present system reduces to keeping only the first term in (44)], nor its generalizations by the addition of gradient terms, have correctly yielded wave-mechanical density oscillations, such as the density oscillations in atoms which correspond to shell structure, or the Friedel oscillations in alloys which are of the same general origin.

<sup>13</sup> J. S. Langer and S. H. Vosko, Phys. Chem. Solids **12**, 196 (1960).

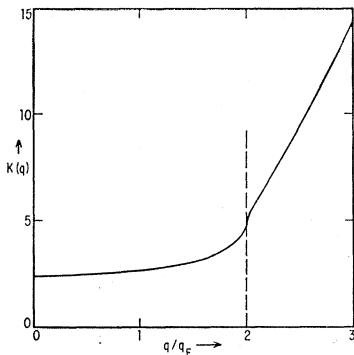


FIG. 2. Behavior of the kernel  $K(q)$ , as a function of  $q$  (electronic density =  $\frac{4}{3} \times 10^{23} \text{ cm}^{-3}$ ).

### III. THE GAS OF SLOWLY VARYING DENSITY

#### 1. The Thomas-Fermi Equation

For a first orientation we shall derive, from our general variational principle, the elementary Thomas-Fermi equation. For this purpose, we use the functional (18) and in (16) we neglect exchange and correlation effects, thus setting  $C_2=0$ . We approximate the kinetic-energy term by its form for a free-electron gas, i.e.,

$$g_r[n] = \frac{3}{10} [k_F(n)]^2 n, \quad (50)$$

where the Fermi momentum  $k_F$  is given by

$$k_F(n) = (3\pi^2 n)^{1/3}. \quad (51)$$

This results in

$$\begin{aligned} E_v[n] &= \int v(\mathbf{r}) n(\mathbf{r}) d\mathbf{r} + \frac{1}{2} \int \frac{n(\mathbf{r}) n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r} d\mathbf{r}' \\ &\quad + \frac{3}{10} (3\pi^2)^{2/3} \int [n(\mathbf{r})]^{5/3} d\mathbf{r}. \end{aligned} \quad (52)$$

To determine  $n(\mathbf{r})$  we now set

$$\delta \left\{ E_v[n] - \mu \int n(\mathbf{r}) d\mathbf{r} \right\} = 0, \quad (53)$$

where  $\mu$  is a Lagrange parameter. This results in the equation

$$v(\mathbf{r}) + \int \frac{n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}' + \frac{1}{2} (3\pi^2)^{2/3} [n(\mathbf{r})]^{2/3} - \mu = 0. \quad (54)$$

If we now introduce the "internal" potential

$$v_i(\mathbf{r}) \equiv \int \frac{n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}', \quad (55)$$

(54) is equivalent to the pair of equations

$$n(\mathbf{r}) = (1/3\pi^2) \{ 2[\mu - v(\mathbf{r}) - v_i(\mathbf{r})] \}^{3/2}, \quad (56)$$

and

$$\nabla^2 v_i(\mathbf{r}) = -4\pi n(\mathbf{r}). \quad (57)$$

From (56) and (57) we can eliminate  $n(\mathbf{r})$  and arrive at

the Thomas-Fermi equation

$$\nabla^2 v_i(\mathbf{r}) = (-2^{7/2}/3\pi) [\mu - v(\mathbf{r}) - v_i(\mathbf{r})]^{3/2}. \quad (58)$$

#### 2. The Gradient Expansion

It is well known that one condition for the validity of the Thomas-Fermi equation is that  $n(\mathbf{r})$  must be a slowly varying function of  $\mathbf{r}$ . This suggests study of the functional  $G[n]$ , where  $n$  has the form

$$n(\mathbf{r}) = \varphi(\mathbf{r}/r_0), \quad (59)$$

with

$$r_0 \rightarrow \infty. \quad (60)$$

It is obvious that this is quite a different class of systems than that considered in Part II ( $n = n_0 + \tilde{n}$ ,  $\tilde{n}/n \ll 1$ ), since now we shall allow  $\varphi$  to have substantial variations. On the other hand, whereas in Part II,  $\tilde{n}$  could contain arbitrarily short wavelengths, these are here ruled out as  $r_0$  becomes large.

We now make the basic assumption that for large  $r_0$ , the partial energy density  $g_r[n]$  may be expanded in the form

$$\begin{aligned} g_r[n] &= g_0(n(\mathbf{r})) + \sum_{i=1}^3 g_i(n(\mathbf{r})) \cdot \nabla_i n(\mathbf{r}) \\ &\quad + \sum_{i,j=1}^3 [g_{i,j}^{(1,1)}(n(\mathbf{r})) \cdot \nabla_i n(\mathbf{r}) \nabla_j n(\mathbf{r}) \\ &\quad \quad + g_{i,j}^{(2)}(n(\mathbf{r})) \cdot \nabla_i \nabla_j n(\mathbf{r})] + \dots \end{aligned} \quad (61)$$

Here successive terms correspond to successive negative powers of the scale parameter  $r_0$ . Quantities like  $g_0(n(\mathbf{r}))$ ,  $g_i(n(\mathbf{r}))$  etc., are functions (not functionals) of  $n(\mathbf{r})$ . No general proof of the existence of such an expansion is known to us, although it can be formally verified in special cases, e.g., when  $G[n(\mathbf{r})]$  can be expanded in powers of  $[n(\mathbf{r}) - n_0]$ . At the same time, we know that, for a finite  $r_0$ , the series does not strictly converge (see the discussion at the end of Sec. II.3), but we may expect it to be useful (in the sense of asymptotic convergence) for sufficiently large values of  $r_0$ .

Now a good deal of progress can be made, using only the fact that  $g_r[n]$  is a universal functional of  $n$ , independent of  $v(\mathbf{r})$ . This requires  $g_r[n]$  to be invariant under rotations about  $\mathbf{r}$ . The coefficients  $g_{i,j}, \dots (n(\mathbf{r}))$ , being functions of the scalar  $n$ , are of course invariant under rotations. Hence one finds by elementary considerations that  $g_r[n]$  must have the form

$$g_r[n] = g_0(n) + [g_2^{(a)}(n) \nabla^2 n + g_2^{(b)}(n) (\nabla n \cdot \nabla n)] + \text{terms of order } \nabla_i^4. \quad (62)$$

A further simplification results from the fact that we may eliminate from  $g_r[n]$  an arbitrary divergence  $\sum_i \nabla_i h_i[n]$  (see the end of Sec. I.3). It is then elementary to show that  $g_r[n]$  may be replaced by

$$\begin{aligned} \tilde{g}_r[n] &= g_0(n) + g_2^{(2)}(n) \nabla n \cdot \nabla n \\ &\quad + \{ g_4^{(2)}(n) (\nabla^2 n) (\nabla^2 n) + g_4^{(3)}(n) (\nabla^2 n) (\nabla n \cdot \nabla n) \\ &\quad \quad + g_4^{(4)}(n) (\nabla n \cdot \nabla n)^2 \} + O(\nabla_i^6). \end{aligned} \quad (63)$$

Here the subscripts refer to the number of gradient operators (or the order in  $1/r_0$ ) and the superscripts to the number of times that  $n$  appears to the right of  $g_\mu^{(\nu)}(n)$ .

It may be worth recalling that while  $\bar{g}_r[n]$  is an admissible density function in the sense that

$$G[n] = \int \bar{g}_r[n] d\mathbf{r}, \quad (64)$$

it differs from the energy density function  $g_r[n]$ , Eq. (18), by a divergence.

### 3. Identification of the Coefficients of the Gradient Expansion

We shall now express the coefficients  $g_\mu^{(\nu)}(n)$  appearing in Eq. (63) in terms of the expansion coefficients, in powers of  $\mathbf{q}$ , of the electronic polarizability  $\alpha(q)$ , and similar higher order, nonlinear, response functions.

We do this by applying our general expression (63) to the case of a nearly uniform electron gas, considered already in Sec. II.2. We go, however, beyond (28) and write

$$n(r) = n_0 + \frac{\lambda}{\Omega} \sum b_1(\mathbf{q}) e^{-i\mathbf{q}\cdot\mathbf{r}} + \frac{\lambda^2}{\Omega} \sum b_2(\mathbf{q}) e^{-i\mathbf{q}\cdot\mathbf{r}} + \dots \quad (65)$$

The linear- and second-, third-, etc., order response functions are then defined by the relations

$$\begin{aligned} b_1(\mathbf{q}) &= \alpha(q) a(\mathbf{q}), \\ b_2(\mathbf{q}) &= \sum_{\mathbf{q}_1 + \mathbf{q}_2 = \mathbf{q}} \alpha(\mathbf{q}_1, \mathbf{q}_2) a(\mathbf{q}_1) a(\mathbf{q}_2), \\ &\text{etc.} \end{aligned} \quad (66)$$

Now let us compare these expressions with what one obtains with the use of (63). We require that

$$\frac{\delta}{\delta n} \left\{ E_v[n] - \mu \int n(\mathbf{r}) d\mathbf{r} \right\} = 0. \quad (67)$$

This gives

$$\begin{aligned} v(\mathbf{r}) + \int \frac{n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}' + g_0' - g_2^{(2)'} (\nabla n)^2 \\ - 2g_2^{(2)} \nabla^2 n + 3g_4^{(2)'} (\nabla^2 n)^2 + 2g_4^{(2)''} (\nabla n)^2 \nabla^2 n \\ + 2g_4^{(2)'} \nabla n \cdot \nabla (\nabla^2 n) + 2g_4^{(2)} (\nabla^2 \nabla^2 n) \\ + g_4^{(3)''} (\nabla n)^4 + 2g_4^{(3)} \nabla n \cdot \nabla (\nabla n)^2 \\ + g_4^{(3)} (\nabla^2 (\nabla n)^2 - 2\nabla n \cdot \nabla (\nabla^2 n) - 2(\nabla^2 n)^2) \\ - 3g_4^{(4)'} (\nabla n)^4 - 4g_4^{(4)} \nabla^2 n (\nabla n)^2 - 4g_4^{(4)} \nabla n \cdot \nabla (\nabla n)^2 \\ + \dots - \mu = 0. \quad (68) \end{aligned}$$

Now let us set

$$v(\mathbf{r}) = \frac{\lambda 4\pi}{\Omega} \sum_{\mathbf{q}} \frac{a(\mathbf{q})}{q^2} e^{-i\mathbf{q}\cdot\mathbf{r}}, \quad (69)$$

$$n = n_0 + \frac{1}{\Omega} \sum_{\mathbf{q}} [\lambda b_1(\mathbf{q}) + \lambda^2 b_2(\mathbf{q}) + \dots] e^{-i\mathbf{q}\cdot\mathbf{r}}, \quad (70)$$

$$\mu = \mu_0 + \lambda \mu_1 + \lambda^2 \mu_2 + \dots \quad (71)$$

Collecting terms of order  $\lambda^0, \lambda^1, \lambda^2$ , we find

$$g_0'(n_0) - \mu_0 = 0, \quad (72)$$

$$\begin{aligned} -\frac{4\pi}{q^2} a(\mathbf{q}) + \left\{ \frac{4\pi}{q^2} + g_0'' + 2g_2^{(2)} q^2 \right. \\ \left. + 2g_4^{(2)} q^4 + \dots \right\} b_1(\mathbf{q}) = 0, \quad (73) \end{aligned}$$

giving

$$b_1(\mathbf{q}) = \left\{ 1 + \left( -\frac{g_0''}{4\pi} \right) q^2 + \left[ \left( \frac{g_0''}{4\pi} \right)^2 - \frac{g_2^{(2)}}{2\pi} \right] q^4 + \dots \right\} a(\mathbf{q}). \quad (74)$$

Also clearly

$$\mu_1 = 0.$$

Similarly, we obtain

$$b_2(\mathbf{q}) = \sum_{\mathbf{q}'} \left\{ \frac{g_0'''}{8\pi} q^2 + \dots \right\} a(\mathbf{q}') a(\mathbf{q} - \mathbf{q}'). \quad (75)$$

If we now expand the response functions in powers of  $q$ ,

$$\alpha(q) = 1 + c_2 q^2 + c_4 q^4 + \dots \quad (76)$$

$$\alpha(\mathbf{q}, \mathbf{q}') = \sum_{m,n} \sum_{i,j} c_{mn}^{ij} q_i^m q_j^n, \quad (77)$$

we can identify the functions  $g_\mu^{(\nu)}$ . Thus

$$g_0''/4\pi = -c_2, \quad (78)$$

$$g_2^{(2)}/4\pi = \frac{1}{2}(-c_4 + c_2^2), \quad (79)$$

$$g_4^{(2)}/4\pi = \frac{1}{2}(-c_6 + 2c_2 c_4 - c_2^3). \quad (80)$$

Similarly all other coefficients  $g_\mu^{(\nu)}(n)$  can be expressed in terms of the expansion coefficients  $c_n$  of the linear polarizability  $\alpha(q)$  of an electron gas of density  $n$ .

In an analogous manner we can express all  $g_\mu^{(\nu)}$  in terms of  $\alpha(\mathbf{q}_1)$  and  $\alpha(\mathbf{q}_1, \mathbf{q}_2)$ ; and generally  $g_\mu^{(\nu)}$  in terms of  $\alpha(\mathbf{q}_1), \dots, \alpha(\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_{\mu-1})$ .

On dimensional grounds we can see from (63) that the gradient expansion requires

$$|\nabla n|/n \ll k_F(n) \quad (81)$$

and

$$|\nabla_i \nabla_j n| / |\nabla n| \ll k_F(n). \quad (82)$$

Both of these conditions are necessary. For while (81) would admit the case of a nearly uniform gas with a small but short-wavelength nonuniformity, this and similar cases are excluded by (82), as they must be.

#### 4. Partial Summation of Gradient Expansion

In the preceding section we have expressed the coefficient  $g_\mu^{(2)}$  in terms of the expansion coefficient  $c_i$  of the polarizability  $\alpha(q)$ , Eq. (76). However, we may apply the expression (63) to the special case of the gas of almost constant density, discussed in Part II. This shows that the leading term  $g_0(n)$  and the subsequent subseries involving coefficients  $g_\mu^{(2)}(n)$  may be summed to yield

$$\begin{aligned} \bar{g}_r[n] = & g_0(n(r)) + \int K_{n(r)}(\mathbf{r}') [n(\mathbf{r} + \frac{1}{2}\mathbf{r}') - n(\mathbf{r})] \\ & \times [n(\mathbf{r} - \frac{1}{2}\mathbf{r}') - n(\mathbf{r})] d\mathbf{r}' + \dots \quad (83) \end{aligned}$$

apart possibly from terms of the form of a divergence or of higher order in the superscript  $\nu$  of  $g_n^{(\nu)}$ . Here

$$K_{n(r)}(\mathbf{r}') = \frac{1}{\Omega} \sum_{\mathbf{q}} \frac{2\pi}{q^2} \left( \frac{1}{\epsilon_{n(r)}(\mathbf{q})} \right) \cdot e^{-i\mathbf{q} \cdot \mathbf{r}'} . \quad (84)$$

The form (83) of  $\bar{g}_r$  has the merit of being exact in both limiting cases where either the density has everywhere nearly the same value (see Part II) or is slowly varying. Its quantitative value for calculating the electronic structure of actual atomic, molecular, or solid-state systems is at present uncertain but is being examined. However, it is already clear that if applied to an atom it will, unlike the simple Thomas-Fermi theory, yield: (1) a finite density at the nucleus, and (2) oscillations in the charge density corresponding to shell structure.

#### 5. Approximate Expressions for the Coefficients of the Gradient Expansion

In the previous section we have expressed the coefficients  $g_\mu^{(\nu)}$  appearing in the gradient expansion (63) in terms of properties of the uniform electron gas. We now collect some results of existing calculations referring to the uniform electron gas which are useful for our present purposes.

$$a. g_0(n)$$

This is the sum of the kinetic+exchange+correlation energy density of a uniform gas of density  $n$ . Here one has available the high-density expansion of Gell-Mann and Brueckner<sup>14</sup>:

$$g_0(n) = \left\{ \frac{2.21}{r_s^2} - \frac{0.916}{r_s} + 0.062 \ln r_s - 0.096 + O(r_s) \right\} n , \quad (85)$$

<sup>14</sup> M. Gell-Mann and K. Brueckner, Phys. Rev. **106**, 364 (1957).

where  $r_s$  is the radius of the Wigner-Seitz sphere defined by

$$\frac{4}{3}\pi r_s^3 = 1/n . \quad (86)$$

This expression is believed to be reasonably accurate only for  $r_s \lesssim 1$ . At lower densities, such as occur in metals ( $2 \lesssim r_s \lesssim 5$ ), various approximate expressions have been proposed. One is due to Wigner<sup>15</sup>

$$g_0(n) \sim \left\{ \frac{2.21}{r_s^2} - \frac{0.916}{r_s} - \frac{0.88}{r_s + 7.8} \right\} n . \quad (87)$$

Other approximations are due to Hubbard,<sup>16</sup> Nozières and Pines,<sup>17</sup> and Gaskell.<sup>18</sup>

$$b. g_\mu^{(2)}(n)$$

These coefficients are all determined in terms of the electronic polarizability,  $\alpha(q)$ . For this latter quantity there is available, at present, a random-phase expression, Eq. (41), which gives

$$\alpha(q) = \frac{2\pi}{k_T^2} \left[ 1 + \frac{q^2}{k_T^2} S(q) \right]^{-1} \quad (88)$$

and

$$\frac{g_2^{(2)}}{4\pi} = \frac{1}{24} \frac{1}{k_T^2 k_F^2} , \quad (89)$$

$$\frac{g_4^{(2)}}{4\pi} = \frac{1}{180} \frac{1}{k_T^2 k_F^4} . \quad (90)$$

Inclusion of the first of these in the energy expression agrees with a correction to the Thomas-Fermi energy functional derived by Kompaneets and Pavlovskii.<sup>3</sup>

An expression for  $\alpha(q)$ , allowing in an approximate manner for exchange effects has been proposed by Hubbard.<sup>16</sup> It is

$$\alpha(q) = \left[ \left( 1 + \frac{1}{2} \frac{q^2}{q^2 + k_F^2} \right) + \frac{q^2}{k_T^2} S(q) \right]^{-1} , \quad (91)$$

where  $S(q)$  is defined in Eq. (43). This form yields

$$\frac{g_2^{(2)}}{4\pi} = \frac{1}{24} \left( \frac{1}{k_T^2 k_F^2} - \frac{6}{k_F^4} \right) . \quad (92)$$

For typical metallic densities this has the opposite sign from the random-phase approximation expression (88). Thus we see that the lowest nonvanishing gradient correction to the Thomas-Fermi theory depends quite sensitively on refinements in the theory of the electronic polarizability,  $\alpha(q)$ .

<sup>15</sup> E. P. Wigner, Phys. Rev. **40**, 1002 (1934).

<sup>16</sup> J. Hubbard, Proc. Roy. Soc. (London) **A243**, 336 (1957).

<sup>17</sup> P. Nozières and D. Pines, Phys. Rev. **111**, 442 (1958).

<sup>18</sup> T. Gaskell, Proc. Phys. Soc. (London) **77**, 1182 (1961); **80**, 1091 (1962).

**IV. CONCLUDING REMARKS**

In the preceding sections we have developed a theory of the electronic ground state which is exact in two limiting cases: The case of a nearly constant density ( $n = n_0 + \bar{n}(r)$ ,  $\bar{n}(r)/n_0 \ll 1$ ) and the case of a slowly varying density. Actual electronic systems do not belong to either of these two categories. The most promising formulation of the theory at present appears to be that obtained by partial summation of the gradient expansion (Sec. III.4). It has, however, not yet been tested in actual physical problems. But regardless of the outcome of this test, it is hoped that the considerations of this paper shed some new light on the problem of the

inhomogeneous electron gas and may suggest further developments.

**ACKNOWLEDGMENTS**

This work was begun and, to a considerable extent, carried out at the University of Paris. One of the authors (P. Hohenberg) acknowledges with thanks a NATO Postdoctoral Fellowship; the other author (W. Kohn) a Guggenheim Fellowship. Both authors wish to thank the faculties of the École Normale Supérieure, Paris, and the Service de Physique des Solides, Orsay, for their hospitality, and Professor A. Blandin, Professor J. Friedel, Dr. R. Balian, and Dr. C. De Dominicis for valuable discussions.

## ON THE EINSTEIN PODOLSKY ROSEN PARADOX\*

J. S. BELL†

*Department of Physics, University of Wisconsin, Madison, Wisconsin*

(Received 4 November 1964)

### I. Introduction

THE paradox of Einstein, Podolsky and Rosen [1] was advanced as an argument that quantum mechanics could not be a complete theory but should be supplemented by additional variables. These additional variables were to restore to the theory causality and locality [2]. In this note that idea will be formulated mathematically and shown to be incompatible with the statistical predictions of quantum mechanics. It is the requirement of locality, or more precisely that the result of a measurement on one system be unaffected by operations on a distant system with which it has interacted in the past, that creates the essential difficulty. There have been attempts [3] to show that even without such a separability or locality requirement no "hidden variable" interpretation of quantum mechanics is possible. These attempts have been examined elsewhere [4] and found wanting. Moreover, a hidden variable interpretation of elementary quantum theory [5] has been explicitly constructed. That particular interpretation has indeed a grossly non-local structure. This is characteristic, according to the result to be proved here, of any such theory which reproduces exactly the quantum mechanical predictions.

### II. Formulation

With the example advocated by Bohm and Aharonov [6], the EPR argument is the following. Consider a pair of spin one-half particles formed somehow in the singlet spin state and moving freely in opposite directions. Measurements can be made, say by Stern-Gerlach magnets, on selected components of the spins  $\vec{\sigma}_1$  and  $\vec{\sigma}_2$ . If measurement of the component  $\vec{\sigma}_1 \cdot \vec{a}$ , where  $\vec{a}$  is some unit vector, yields the value +1 then, according to quantum mechanics, measurement of  $\vec{\sigma}_2 \cdot \vec{a}$  must yield the value -1 and vice versa. Now we make the hypothesis [2], and it seems one at least worth considering, that if the two measurements are made at places remote from one another the orientation of one magnet does not influence the result obtained with the other. Since we can predict in advance the result of measuring any chosen component of  $\vec{\sigma}_2$ , by previously measuring the same component of  $\vec{\sigma}_1$ , it follows that the result of any such measurement must actually be predetermined. Since the initial quantum mechanical wave function does not determine the result of an individual measurement, this predetermination implies the possibility of a more complete specification of the state.

Let this more complete specification be effected by means of parameters  $\lambda$ . It is a matter of indifference in the following whether  $\lambda$  denotes a single variable or a set, or even a set of functions, and whether the variables are discrete or continuous. However, we write as if  $\lambda$  were a single continuous parameter. The result  $A$  of measuring  $\vec{\sigma}_1 \cdot \vec{a}$  is then determined by  $\vec{a}$  and  $\lambda$ , and the result  $B$  of measuring  $\vec{\sigma}_2 \cdot \vec{b}$  in the same instance is determined by  $\vec{b}$  and  $\lambda$ , and

\*Work supported in part by the U.S. Atomic Energy Commission

†On leave of absence from SLAC and CERN

$$A(\vec{a}, \lambda) = \pm 1, B(\vec{b}, \lambda) = \pm 1. \quad (1)$$

The vital assumption [2] is that the result  $B$  for particle 2 does not depend on the setting  $\vec{a}$ , of the magnet for particle 1, nor  $A$  on  $\vec{b}$ .

If  $\rho(\lambda)$  is the probability distribution of  $\lambda$  then the expectation value of the product of the two components  $\vec{\sigma}_1 \cdot \vec{a}$  and  $\vec{\sigma}_2 \cdot \vec{b}$  is

$$P(\vec{a}, \vec{b}) = \int d\lambda \rho(\lambda) A(\vec{a}, \lambda) B(\vec{b}, \lambda) \quad (2)$$

This should equal the quantum mechanical expectation value, which for the singlet state is

$$\langle \vec{\sigma}_1 \cdot \vec{a} \vec{\sigma}_2 \cdot \vec{b} \rangle = -\vec{a} \cdot \vec{b}. \quad (3)$$

But it will be shown that this is not possible.

Some might prefer a formulation in which the hidden variables fall into two sets, with  $A$  dependent on one and  $B$  on the other; this possibility is contained in the above, since  $\lambda$  stands for any number of variables and the dependences thereon of  $A$  and  $B$  are unrestricted. In a complete physical theory of the type envisaged by Einstein, the hidden variables would have dynamical significance and laws of motion; our  $\lambda$  can then be thought of as initial values of these variables at some suitable instant.

### III. Illustration

The proof of the main result is quite simple. Before giving it, however, a number of illustrations may serve to put it in perspective.

Firstly, there is no difficulty in giving a hidden variable account of spin measurements on a single particle. Suppose we have a spin half particle in a pure spin state with polarization denoted by a unit vector  $\vec{p}$ . Let the hidden variable be (for example) a unit vector  $\vec{\lambda}$  with uniform probability distribution over the hemisphere  $\vec{\lambda} \cdot \vec{p} > 0$ . Specify that the result of measurement of a component  $\vec{\sigma} \cdot \vec{a}$  is

$$\text{sign } \vec{\lambda} \cdot \vec{a}', \quad (4)$$

where  $\vec{a}'$  is a unit vector depending on  $\vec{a}$  and  $\vec{p}$  in a way to be specified, and the sign function is +1 or -1 according to the sign of its argument. Actually this leaves the result undetermined when  $\lambda \cdot a' = 0$ , but as the probability of this is zero we will not make special prescriptions for it. Averaging over  $\vec{\lambda}$  the expectation value is

$$\langle \vec{\sigma} \cdot \vec{a} \rangle = 1 - 2\theta'/\pi, \quad (5)$$

where  $\theta'$  is the angle between  $\vec{a}'$  and  $\vec{p}$ . Suppose then that  $\vec{a}'$  is obtained from  $\vec{a}$  by rotation towards  $\vec{p}$  until

$$1 - \frac{2\theta'}{\pi} = \cos \theta \quad (6)$$

where  $\theta$  is the angle between  $\vec{a}$  and  $\vec{p}$ . Then we have the desired result

$$\langle \vec{\sigma} \cdot \vec{a} \rangle = \cos \theta \quad (7)$$

So in this simple case there is no difficulty in the view that the result of every measurement is determined by the value of an extra variable, and that the statistical features of quantum mechanics arise because the value of this variable is unknown in individual instances.

Secondly, there is no difficulty in reproducing, in the form (2), the only features of (3) commonly used in verbal discussions of this problem:

$$\left. \begin{aligned} P(\vec{a}, \vec{a}) &= -P(\vec{a}, -\vec{a}) = -1 \\ P(\vec{a}, \vec{b}) &= 0 \text{ if } \vec{a} \cdot \vec{b} = 0 \end{aligned} \right\} \quad (8)$$

For example, let  $\lambda$  now be unit vector  $\vec{\lambda}$ , with uniform probability distribution over all directions, and take

$$\left. \begin{aligned} A(\vec{a}, \vec{\lambda}) &= \text{sign } \vec{a} \cdot \vec{\lambda} \\ B(\vec{a}, \vec{\lambda}) &= -\text{sign } \vec{b} \cdot \vec{\lambda} \end{aligned} \right\} \quad (9)$$

This gives

$$P(\vec{a}, \vec{b}) = -1 + \frac{2}{\pi} \theta, \quad (10)$$

where  $\theta$  is the angle between  $a$  and  $b$ , and (10) has the properties (8). For comparison, consider the result of a modified theory [6] in which the pure singlet state is replaced in the course of time by an isotropic mixture of product states; this gives the correlation function

$$- \frac{1}{3} \vec{a} \cdot \vec{b} \quad (11)$$

It is probably less easy, experimentally, to distinguish (10) from (3), than (11) from (3).

Unlike (3), the function (10) is not stationary at the minimum value  $-1$  (at  $\theta = 0$ ). It will be seen that this is characteristic of functions of type (2).

Thirdly, and finally, there is no difficulty in reproducing the quantum mechanical correlation (3) if the results  $A$  and  $B$  in (2) are allowed to depend on  $\vec{b}$  and  $\vec{a}$  respectively as well as on  $\vec{a}$  and  $\vec{b}$ . For example, replace  $\vec{a}$  in (9) by  $\vec{a}'$ , obtained from  $\vec{a}$  by rotation towards  $\vec{b}$  until

$$1 - \frac{2}{\pi} \theta' = \cos \theta,$$

where  $\theta'$  is the angle between  $\vec{a}'$  and  $\vec{b}$ . However, for given values of the hidden variables, the results of measurements with one magnet now depend on the setting of the distant magnet, which is just what we would wish to avoid.

#### IV. Contradiction

The main result will now be proved. Because  $\rho$  is a normalized probability distribution,

$$\int d\lambda \rho(\lambda) = 1, \quad (12)$$

and because of the properties (1),  $P$  in (2) cannot be less than  $-1$ . It can reach  $-1$  at  $\vec{a} = \vec{b}$  only if

$$A(\vec{a}, \lambda) = -B(\vec{a}, \lambda) \quad (13)$$

except at a set of points  $\lambda$  of zero probability. Assuming this, (2) can be rewritten

$$P(\vec{a}, \vec{b}) = - \int d\lambda \rho(\lambda) A(\vec{a}, \lambda) A(\vec{b}, \lambda). \quad (14)$$

It follows that  $\vec{c}$  is another unit vector

$$\begin{aligned} P(\vec{a}, \vec{b}) - P(\vec{a}, \vec{c}) &= - \int d\lambda \rho(\lambda) [A(\vec{a}, \lambda) A(\vec{b}, \lambda) - A(\vec{a}, \lambda) A(\vec{c}, \lambda)] \\ &= \int d\lambda \rho(\lambda) A(\vec{a}, \lambda) A(\vec{b}, \lambda) [A(\vec{b}, \lambda) A(\vec{c}, \lambda) - 1] \end{aligned}$$

using (1), whence

$$|P(\vec{a}, \vec{b}) - P(\vec{a}, \vec{c})| \leq \int d\lambda \rho(\lambda) [1 - A(\vec{b}, \lambda) A(\vec{c}, \lambda)]$$

The second term on the right is  $P(\vec{b}, \vec{c})$ , whence

$$1 + P(\vec{b}, \vec{c}) \geq |P(\vec{a}, \vec{b}) - P(\vec{a}, \vec{c})| \quad (15)$$

Unless  $P$  is constant, the right hand side is in general of order  $|\vec{b} - \vec{c}|$  for small  $|\vec{b} - \vec{c}|$ . Thus  $P(\vec{b}, \vec{c})$  cannot be stationary at the minimum value (-1 at  $\vec{b} = \vec{c}$ ) and cannot equal the quantum mechanical value (3).

Nor can the quantum mechanical correlation (3) be arbitrarily closely approximated by the form (2). The formal proof of this may be set out as follows. We would not worry about failure of the approximation at isolated points, so let us consider instead of (2) and (3) the functions

$$\bar{P}(\vec{a}, \vec{b}) \text{ and } \overline{-\vec{a} \cdot \vec{b}}$$

where the bar denotes independent averaging of  $P(\vec{a}', \vec{b}')$  and  $-\vec{a}' \cdot \vec{b}'$  over vectors  $\vec{a}'$  and  $\vec{b}'$  within specified small angles of  $\vec{a}$  and  $\vec{b}$ . Suppose that for all  $\vec{a}$  and  $\vec{b}$  the difference is bounded by  $\epsilon$ :

$$|\bar{P}(\vec{a}, \vec{b}) + \vec{a} \cdot \vec{b}| \leq \epsilon \quad (16)$$

Then it will be shown that  $\epsilon$  cannot be made arbitrarily small.

Suppose that for all  $a$  and  $b$

$$|\overline{\vec{a} \cdot \vec{b}} - \vec{a} \cdot \vec{b}| \leq \delta \quad (17)$$

Then from (16)

$$|\bar{P}(\vec{a}, \vec{b}) + \vec{a} \cdot \vec{b}| \leq \epsilon + \delta \quad (18)$$

From (2)

$$\bar{P}(\vec{a}, \vec{b}) = \int d\lambda \rho(\lambda) \bar{A}(\vec{a}, \lambda) \bar{B}(\vec{b}, \lambda) \quad (19)$$

where

$$|\bar{A}(\vec{a}, \lambda)| \leq 1 \text{ and } |\bar{B}(\vec{b}, \lambda)| \leq 1 \quad (20)$$

From (18) and (19), with  $\vec{a} = \vec{b}$ ,

$$d\lambda \rho(\lambda) [\bar{A}(\vec{b}, \lambda) \bar{B}(\vec{b}, \lambda) + 1] \leq \epsilon + \delta \quad (21)$$

From (19)

$$\begin{aligned} \bar{P}(\vec{a}, \vec{b}) - \bar{P}(\vec{a}, \vec{c}) &= \int d\lambda \rho(\lambda) [\bar{A}(\vec{a}, \lambda) \bar{B}(\vec{b}, \lambda) - \bar{A}(\vec{a}, \lambda) \bar{B}(\vec{c}, \lambda)] \\ &= \int d\lambda \rho(\lambda) \bar{A}(\vec{a}, \lambda) \bar{B}(\vec{b}, \lambda) [1 + \bar{A}(\vec{b}, \lambda) \bar{B}(\vec{c}, \lambda)] \\ &\quad - \int d\lambda \rho(\lambda) \bar{A}(\vec{a}, \lambda) \bar{B}(\vec{c}, \lambda) [1 + \bar{A}(\vec{b}, \lambda) \bar{B}(\vec{b}, \lambda)] \end{aligned}$$

Using (20) then

$$|\bar{P}(\vec{a}, \vec{b}) - \bar{P}(\vec{a}, \vec{c})| \leq \int d\lambda \alpha(\lambda) [1 + \bar{A}(\vec{b}, \lambda) \bar{B}(\vec{c}, \lambda)] \\ + \int d\lambda \rho(\lambda) [1 + \bar{A}(\vec{b}, \lambda) \bar{B}(\vec{b}, \lambda)]$$

Then using (19) and 21)

$$|\bar{P}(\vec{a}, \vec{b}) - \bar{P}(\vec{a}, \vec{c})| \leq 1 + \bar{P}(\vec{b}, \vec{c}) + \epsilon + \delta$$

Finally, using (18),

$$|\vec{a} \cdot \vec{c} - \vec{a} \cdot \vec{b}| - 2(\epsilon + \delta) \leq 1 - \vec{b} \cdot \vec{c} + 2(\epsilon + \delta)$$

or

$$4(\epsilon + \delta) \geq |\vec{a} \cdot \vec{c} - \vec{a} \cdot \vec{b}| + \vec{b} \cdot \vec{c} - 1 \quad (22)$$

Take for example  $\vec{a} \cdot \vec{c} = 0$ ,  $\vec{a} \cdot \vec{b} = \vec{b} \cdot \vec{c} = 1/\sqrt{2}$  Then

$$4(\epsilon + \delta) \geq \sqrt{2} - 1$$

Therefore, for small finite  $\delta$ ;  $\epsilon$  cannot be arbitrarily small.

Thus, the quantum mechanical expectation value cannot be represented, either accurately or arbitrarily closely, in the form (2).

## V. Generalization

The example considered above has the advantage that it requires little imagination to envisage the measurements involved actually being made. In a more formal way, assuming [7] that any Hermitian operator with a complete set of eigenstates is an "observable", the result is easily extended to other systems. If the two systems have state spaces of dimensionality greater than 2 we can always consider two dimensional subspaces and define, in their direct product, operators  $\vec{\sigma}_1$  and  $\vec{\sigma}_2$  formally analogous to those used above and which are zero for states outside the product subspace. Then for at least one quantum mechanical state, the "singlet" state in the combined subspaces, the statistical predictions of quantum mechanics are incompatible with separable predetermination.

## VI. Conclusion

In a theory in which parameters are added to quantum mechanics to determine the results of individual measurements, without changing the statistical predictions, there must be a mechanism whereby the setting of one measuring device can influence the reading of another instrument, however remote. Moreover, the signal involved must propagate instantaneously, so that such a theory could not be Lorentz invariant.

Of course, the situation is different if the quantum mechanical predictions are of limited validity. Conceivably they might apply only to experiments in which the settings of the instruments are made sufficiently in advance to allow them to reach some mutual rapport by exchange of signals with velocity less than or equal to that of light. In that connection, experiments of the type proposed by Bohm and Aharonov [6], in which the settings are changed during the flight of the particles, are crucial.

*I am indebted to Drs. M. Bander and J. K. Perring for very useful discussions of this problem. The first draft of the paper was written during a stay at Brandeis University; I am indebted to colleagues there and at the University of Wisconsin for their interest and hospitality.*

### References

1. A. EINSTEIN, N. ROSEN and B. PODOLSKY, *Phys. Rev.* **47**, 777 (1935); see also N. BOHR, *Ibid.* **48**, 696 (1935), W. H. FURRY, *Ibid.* **49**, 393 and 476 (1936), and D. R. INGLIS, *Rev. Mod. Phys.* **33**, 1 (1961).
2. "But on one supposition we should, in my opinion, absolutely hold fast: the real factual situation of the system  $S_2$  is independent of what is done with the system  $S_1$ , which is spatially separated from the former." A. EINSTEIN in *Albert Einstein, Philosopher Scientist*, (Edited by P. A. SCHILP) p. 85, Library of Living Philosophers, Evanston, Illinois (1949).
3. J. VON NEUMANN, *Mathematische Grundlagen der Quanten-mechanik*. Verlag Julius-Springer, Berlin (1932), [English translation: Princeton University Press (1955)]; J. M. JAUCH and C. PIRON, *Helv. Phys. Acta* **36**, 827 (1963).
4. J. S. BELL, to be published.
5. D. BOHM, *Phys. Rev.* **85**, 166 and 180 (1952).
6. D. BOHM and Y. AHARONOV, *Phys. Rev.* **108**, 1070 (1957).
7. P. A. M. DIRAC, *The Principles of Quantum Mechanics* (3rd Ed.) p. 37. The Clarendon Press, Oxford (1947).

Three-Triplet Model with Double  $SU(3)$  Symmetry\*

M. Y. HAN

*Department of Physics, Syracuse University, Syracuse, New York*

AND

Y. NAMBU

*The Enrico Fermi Institute for Nuclear Studies, and the Department of Physics,  
The University of Chicago, Chicago, Illinois*

(Received 12 April 1965)

With a view to avoiding some of the kinematical and dynamical difficulties involved in the single-triplet quark model, a model for the low-lying baryons and mesons based on three triplets with integral charges is proposed, somewhat similar to the two-triplet model introduced earlier by one of us (Y. N.). It is shown that in a  $U(3)$  scheme of triplets with integral charges, one is naturally led to three triplets located symmetrically about the origin of  $I_3-Y$  diagram under the constraint that the Nishijima-Gell-Mann relation remains intact. A double  $SU(3)$  symmetry scheme is proposed in which the large mass splittings between different representations are ascribed to one of the  $SU(3)$ , while the other  $SU(3)$  is the usual one for the mass splittings within a representation of the first  $SU(3)$ .

## I. INTRODUCTION

**A**LTHOUGH the  $SU(6)$  symmetry strongly indicates that the baryon is essentially a three-body system built from some basic triplet field or fields, the quark model<sup>1</sup> is not entirely satisfactory from a realistic point of view, because (a) the electric charges are not integral, (b) three quarks in  $s$  states do not form the symmetric  $SU(6)$  representation assigned to the baryons, and (c) a simple dynamical mechanism is lacking for realizing only zero-triality states as the low-lying levels.

These difficulties may be avoided if we introduce more than one basic triplet. Recently one of us (Y. N.) has attempted a two-triplet model<sup>2</sup> where the members of the triplets  $t_1$  and  $t_2$  had the charge assignment  $(1,0,0)$  and  $(0,-1,-1)$ , as had been proposed earlier by Bacry *et al.*<sup>3</sup> The baryon would be represented by the combination  $t_1 t_1 t_2$ , whereas the mesons would correspond to some combination  $\sim a t_1 t_1' + b t_2 t_2'$ . The triplets are assumed to have masses large compared to the baryon mass, which would mean that baryons and mesons have very large binding energies. A dynamical mechanism for this is provided by a neutral field coupled strongly to the "charm number"<sup>4</sup>  $C$ , which is 1 for  $t_1$  and -2 for  $t_2$ , and therefore  $C=0$  for baryons and mesons. In analogy with electrostatic energy, we can argue that the potential energy due to the charm field would be lowest when the system is "neutral," namely,  $C=0$ . Thus all

other unwanted configurations with  $C \neq 0$ , which include among others triplet, sextet, etc. representations, would have high masses, and hence would not be easily observed.

There have been proposed two different ways in which to introduce basic triplet or triplets with integral charges. One approach essentially involves a modification of the Nishijima-Gell-Mann relation by way of introducing an additional quantum number, the triality quantum number,<sup>5</sup> and this has led to considerations of higher symmetry schemes based on rank-three Lie groups.<sup>6</sup> On the other hand, Okubo *et al.*<sup>7</sup> have recently shown that the minimal group required for this purpose is actually the group  $U(3)$ .<sup>8</sup> It is shown that a triplet scheme may be defined in  $U(3)$  such that the triplet always possesses integral values of charge and hypercharge and satisfies the Nishijima-Gell-Mann relation without a modification. The  $U(3)$  triplet considered by Okubo *et al.* is of Sakata type; i.e., it consists of an isodoublet and an isosinglet. Actually, the  $U(3)$  scheme is much more appealing than those of the rank-three Lie groups on two accounts: firstly, the Nishijima-Gell-Mann relation is satisfied universally by triplets as by octets and decuplets, and secondly as far as the hitherto realized representations are concerned,  $U(3)$  is equivalent to  $SU(3)$ .<sup>9</sup>

In what follows, we show that the  $U(3)$  scheme, when fully utilized as described below, naturally and uniquely

\* Work supported in part by the U. S. Atomic Energy Commission under the Contract No. AT(30-1)-3399 and No. AT(11-1)-264.

<sup>1</sup> M. Gell-Mann, Phys. Letters 8, 214 (1964); G. Zweig, CERN (to be published).

<sup>2</sup> Y. Nambu, *Proceedings of the Second Coral Gables Conference on Symmetry Principles at High Energy* (W. H. Freeman and Company, San Francisco, 1965).

<sup>3</sup> H. Bacry, J. Nuyts, and L. van Hove, Phys. Letters 9, 279 (1964).

<sup>4</sup> This name was originally used in connection with the  $SU(4)$  symmetry. B. J. Bjørken and S. L. Glashow, Phys. Letters 11, 255 (1964); A. Salam, Dubna Conference Report, 1964 (unpublished).

<sup>5</sup> G. E. Baird and L. C. Biedenharn, *Proceedings of the First Coral Gables Conference on Symmetry Principles at High Energy* (W. H. Freeman and Company, San Francisco, 1964); C. R. Hagen and A. J. Macfarlane, Phys. Rev. 135, B432 (1964) and J. Math. Phys. 5, 1335 (1964).

<sup>6</sup> For example, see I. S. Gerstein and M. L. Whippmann, Phys. Rev. 137, B1522 (1965). Earlier references are given in this paper.

<sup>7</sup> S. Okubo, C. Ryan, and R. E. Marshak, Nuovo Cimento 34, 759 (1964).

<sup>8</sup> The use of  $U(3)$  in this connection has also been remarked by I. S. Gerstein and K. T. Mahanthappa, Phys. Rev. Letters 12, 570, 656(E) (1964).

<sup>9</sup> S. Okubo, Phys. Letters 4, 14 (1963).

leads to a set of three basic triplets with integral charges, namely an  $I$ -triplet (isodoublet and isosinglet), a  $U$ -triplet ( $U$ -spin doublet and  $U$ -spin singlet) and a  $V$ -triplet ( $V$ -spin doublet and  $V$ -spin singlet).<sup>10</sup> These triplets arise from three different ways of defining charge  $Q$ , hypercharge  $Y$ , and a displaced isospin  $I_3$  in the  $U(3)$  group as opposed to the  $SU(3)$ , in such a way that the charge and hypercharge have integral values, while keeping the Nishijima-Gell-Mann relation intact, and they differ from each other in their quantum-number assignments as well as in their transformation properties under the Weyl reflections.<sup>11</sup> This is described in Sec. II. In Sec. III, a double  $SU(3)$  symmetry scheme is proposed based on the three-triplet model in which the large mass splittings between different representations are ascribed to one of the  $SU(3)$ , and the other  $SU(3)$  is, as usual, responsible for the mass splittings within a representation. The low-lying baryon and meson states may be taken as singlets with respect to one of the  $SU(3)$ . The extended symmetry group with respect to the  $SU(6)$  symmetry is briefly discussed.

## II. THREE TRIPLETS

We shall denote the infinitesimal generators of  $U(3)$  by  $A_\nu^\mu$  which satisfies the following commutation relations:

$$[A_\beta^\alpha, A_\nu^\mu] = \delta_\beta^\mu A_\nu^\alpha - \delta_\nu^\alpha A_\beta^\mu, \quad (1)$$

where all indices take on the values 1, 2, and 3. The corresponding infinitesimal generators  $B_\nu^\mu$  of  $SU(3)$  are then given by

$$B_\nu^\mu = A_\nu^\mu - \frac{1}{3} \delta_\nu^\mu A_\lambda^\lambda \quad (2)$$

which satisfy the following equations:

$$[B_\beta^\alpha, B_\nu^\mu] = \delta_\beta^\mu B_\nu^\alpha - \delta_\nu^\alpha B_\beta^\mu \quad (3)$$

and

$$B_\lambda^\lambda = 0. \quad (4)$$

Furthermore, the unitary restriction gives

$$(A_\mu^\mu)^\dagger = A_\mu^\mu, \quad (B_\mu^\mu)^\dagger = B_\mu^\mu. \quad (5)$$

Let us now briefly summarize the relevant results of Okubo *et al.* In the  $SU(3)$  scheme, the charge  $Q$ , the hypercharge  $Y$  and the third component of isospin  $I_3$  are identified as follows<sup>12</sup>:

$$Q = -B_1^1, \quad (6a)$$

$$Y = B_3^3 = -B_1^1 - B_2^2 \quad [\text{by the relation (4)}], \quad (6b)$$

$$I_3 = \frac{1}{2}(B_2^2 - B_1^1). \quad (6c)$$

In the  $U(3)$  scheme, the corresponding quantities  $\tilde{Q}$ ,  $\tilde{Y}$ ,

<sup>10</sup> C. A. Levinson, H. J. Lipkin, and S. Meshkov, Nuovo Cimento **23**, 236 (1961); Phys. Letters **1**, 44 (1962) and Phys. Rev. Letters **10**, 361 (1963).

<sup>11</sup> A. J. Macfarlane, E. C. G. Sudarshan, and C. Dullemond, Nuovo Cimento **30**, 845 (1963).

<sup>12</sup> We use the sign convention of S. P. Rosen, J. Math. Phys. **5**, 289 (1964).

and  $\tilde{I}_3$  are defined as follows:

$$\tilde{Q} = -A_1^1 = Q - \frac{1}{3}\tau, \quad (7a)$$

$$\tilde{Y} = -A_2^2 - A_3^3 = Y - \frac{2}{3}\tau, \quad (7b)$$

$$\tilde{I}_3 = \frac{1}{2}(A_2^2 - A_1^1) = I_3, \quad (7c)$$

where

$$\tau = A_1^1 + A_2^2 + A_3^3. \quad (8)$$

With these definitions, the Nishijima-Gell-Mann relation is seen to be equally satisfied by the  $U(3)$  and  $SU(3)$  theories, i.e.,

$$Q = I_3 + \frac{1}{2}Y \quad (9)$$

and

$$\tilde{Q} = \tilde{I}_3 + \frac{1}{2}\tilde{Y}, \quad (10)$$

respectively. Since the generators  $A_1^1$ ,  $A_2^2$ , and  $A_3^3$  possess integral eigenvalues in any representation,<sup>13</sup> the identifications of  $\tilde{Q}$  and  $\tilde{Y}$  to be the charge and the hypercharge, respectively, in  $U(3)$  theory shall always lead to integral values for the charge and the hypercharge. In particular, in the three-dimensional representation, the  $U(3)$  triplet has the eigenvalues

$$Q = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \tilde{I}_3 = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & -\frac{1}{2} & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \tilde{Y} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \quad (11)$$

This triplet corresponds to the Sakata triplet which we call an  $I$  triplet for short.

We can now generalize the above constructions of the  $U(3)$  triplet in the following way. Comparing (6b) and (7b), we see that a particular choice has been made for  $\tilde{Y}$ . Had we defined  $\tilde{Y}$  to be  $A_3^3$ , it would still have integral eigenvalues but the relation (10) would have been violated. This is because  $B_\lambda^\lambda = 0$  in  $SU(3)$  but  $A_\lambda^\lambda \neq 0$  in general in  $U(3)$  and thus some care is needed in defining corresponding quantities in  $U(3)$ . Making use of (4), the definition in (6) can be written more generally as

$$Q = -B_1^1 = B_2^2 + B_3^3, \quad (12a)$$

$$Y = B_3^3 = -B_1^1 - B_2^2, \quad (12b)$$

$$I_3 = \frac{1}{2}(B_2^2 - B_1^1) = \frac{1}{2}(2B_2^2 + B_3^3) = -\frac{1}{2}(2B_1^1 + B_3^3). \quad (12c)$$

As in (7), replacing  $B_\nu^\mu$ 's in (12) by corresponding  $A_\nu^\mu$ 's, we list all possible candidates for the corresponding quantities in  $U(3)$  which are now however not equivalent to each other [they are equivalent, of course, when reduced to  $SU(3)$ ], i.e.,

$$\tilde{Q}: -A_1^1, \quad A_2^2 + A_3^3, \quad (13a)$$

$$\tilde{Y}: A_3^3, \quad -A_1^1 - A_2^2, \quad (13b)$$

$$\tilde{I}_3: \frac{1}{2}(A_2^2 - A_1^1), \quad \frac{1}{2}(2A_2^2 + A_3^3), \quad -\frac{1}{2}(2A_1^1 + A_3^3). \quad (13c)$$

<sup>13</sup> For a derivation of this result, see Eq. (7) of Ref. 7.

To start with, the alternative choices in (13) provide twelve inequivalent ways in which to choose a set of three quantities  $\tilde{Q}$ ,  $\tilde{Y}$  and  $\tilde{I}_3$  for the  $U(3)$  scheme. In every choice  $\tilde{Q}$  and  $\tilde{Y}$  will have integral eigenvalues, but as can be easily checked the Nishijima-Gell-Mann relation will not be valid for all of them. In fact, there are only three cases for which it is valid and we are thus naturally led to three inequivalent triplets in the  $U(3)$  scheme; they are defined by the following three choices:

$$t_I: \quad \tilde{Q} = -A_1^1, \quad \tilde{Y} = -A_1^1 - A_2^2, \quad \tilde{I}_3 = \frac{1}{2}(A_2^2 - A_1^1), \quad (14a)$$

$$t_U: \quad \tilde{Q} = A_2^2 + A_3^3, \quad \tilde{Y} = A_3^3, \quad \tilde{I}_3 = \frac{1}{2}(2A_2^2 + A_3^3), \quad (14b)$$

$$t_V: \quad \tilde{Q} = -A_1^1, \quad \tilde{Y} = A_3^3, \quad \tilde{I}_3 = -\frac{1}{2}(2A_1^1 + A_3^3). \quad (14c)$$

Now the first one,  $t_I$ , for which

$$\tilde{Y} = -A_1^1 - A_2^2, \quad (15)$$

$$\tilde{I}_3 = \frac{1}{2}(A_2^2 - A_1^1) = \frac{1}{2}(B_2^2 - B_1^1) = I_3 \quad (16)$$

corresponds to the  $I$  triplet mentioned above.

The structure of the remaining triplets  $t_U$  and  $t_V$  can be brought to much more transparent and symmetric forms in terms of the  $U$ -spin and  $V$ -spin subalgebras.<sup>10</sup> As in the case of relations (9) and (10) for  $SU(3)$  and  $U(3)$ , we define the  $U$  and  $V$  spin of  $U(3)$  in exactly the same forms as in  $SU(3)$  except that all quantities are tilded quantities. From the  $SU(3)$  definitions,<sup>12</sup> we then have

$$\tilde{Y}_U = -\tilde{Q} = -A_2^2 - A_3^3, \quad (17)$$

$$\tilde{U}_3 = \tilde{Y} - \frac{1}{2}\tilde{Q} = \frac{1}{2}(A_3^3 - A_2^2) = \frac{1}{2}(B_3^3 - B_2^2) = U_3 \quad (18)$$

for (14b), and

$$\tilde{Y}_V = \tilde{Q} - \tilde{Y} = -A_3^3 - A_1^1, \quad (19)$$

$$\tilde{V}_3 = -\frac{1}{2}(\tilde{Y} + \tilde{Q}) = \frac{1}{2}(A_1^1 - A_3^3) = \frac{1}{2}(B_1^1 - B_3^3) = V_3 \quad (20)$$

for (14c). They correspond, therefore, to a  $U$  triplet and a  $V$  triplet, respectively, and hence the notations  $t_I$ ,  $t_U$ , and  $t_V$ . With respect to the  $SU(3)$  triplet (quark), these  $U(3)$  triplets have their respective "hypercharges" (i.e.,  $Y$ ,  $Y_U$ , and  $Y_V$ ) shifted by the amount of  $\frac{2}{3}$  and as such they have quite different transformation properties under the Weyl reflections  $W_1$ ,  $W_2$ , and  $W_3$ <sup>11</sup> which are reflections about the axis  $I_3=0$ ,  $U_3=0$ , and  $V_3=0$ , respectively. Whereas the  $SU(3)$  triplet is invariant under all three Weyl reflections, the  $U(3)$  triplets are not. They transform according to

$$W_1: \quad t_I \rightarrow t_I, \quad t_U \leftrightarrow t_V; \quad (21a)$$

$$W_2: \quad t_U \rightarrow t_U, \quad t_I \leftrightarrow t_V; \quad (21b)$$

$$W_3: \quad t_V \rightarrow t_V, \quad t_I \leftrightarrow t_U. \quad (21c)$$

Figure 1 and Table I(a) list the quantum numbers  $\tilde{I}_3$  and  $\tilde{Y}$  for the single triplet (quark) model; a possible

TABLE I. Quantum-number assignments for (a) the quark model, (b) the two-triplet model, and (c) the three-triplet model.

(a) quark						
$\tilde{I}_3$	$\frac{1}{2}$	$-\frac{1}{2}$	0	$t_1$	$t_2$	$t_3$
$\tilde{Y}$	$\frac{1}{3}$	$\frac{1}{3}$	$-\frac{2}{3}$			
$\tilde{Q}$	$\frac{2}{3}$	$-\frac{1}{3}$	$-\frac{1}{3}$			

(b)						
$\tilde{I}_3$	$\frac{1}{2}$	$-\frac{1}{2}$	0	$t_1$	$t_2$	$t_3$
$\tilde{Y}$	1	1	0		$-\frac{1}{2}$	$-2$
$\tilde{Q}$	1	0	0		0	-1

(c)						
$\tilde{I}_3$	$\frac{1}{2}$	$-\frac{1}{2}$	0	$t_1(t_I)$	$t_2(t_U)$	$t_3(t_V)$
$\tilde{Y}$	1	1	0		0	$\frac{1}{2}$
$\tilde{Q}$	1	0	0		0	0

assignment implied by the two-triplet model<sup>2</sup> is shown in Fig. 2 and Table I(b); the corresponding quantum numbers for the three-triplet model are given in Fig. 3 and Table I(c).

### III. DOUBLE $SU(3)$ SYMMETRY

Let us call the three triplets  $t_1 (= t_I)$ ,  $t_2 (= t_U)$ , and  $t_3 (= t_V)$ . Each triplet may be characterized in general by the average values,  $\tilde{I}_3$  and  $\tilde{Y}$ , of  $\tilde{I}_3$  and  $\tilde{Y}$  for its three members. This specifies the location of the center of the triplet in the  $\tilde{I}_3$ - $\tilde{Y}$  diagram. Since  $\tilde{A}_1^1 = \tilde{A}_2^2 = \tilde{A}_3^3 = \bar{\tau}/3 = \tau/3$ , Eq. (14) gives for the three definitions of

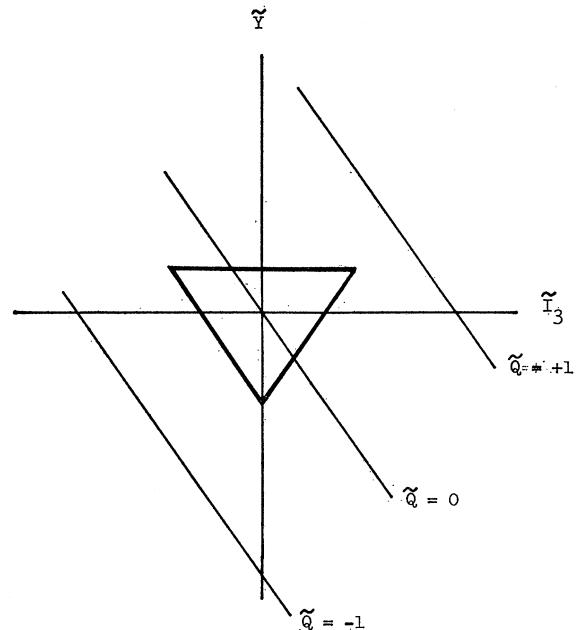


FIG. 1. The single-triplet (quark) model.

$\tilde{I}_3$  and  $\tilde{Y}$ ,

$$\begin{aligned}\tilde{I}_3 &= 0, \frac{1}{2}\tau, -\frac{1}{2}\tau, \\ \tilde{Y} &= -\frac{2}{3}\tau, \frac{1}{3}\tau, \frac{1}{3}\tau,\end{aligned}\quad (22)$$

respectively, where  $\tau = -1$  for all the triplets. We may define new quantities  $I_3$ ,  $Y$  and  $Q = I_3 + \frac{1}{2}Y$  by the relations:

$$\begin{aligned}\tilde{I}_3 &= \tilde{I}_3 + I_3, \\ \tilde{Y} &= \tilde{Y} + Y, \\ \tilde{Q} &= \tilde{I}_3 + \frac{1}{2}\tilde{Y} + I_3 + \frac{1}{2}Y = Q + Q.\end{aligned}\quad (23)$$

It is clear that  $I_3$  and  $Y$  play the role of  $SU(3)$  generators within each triplet. The charm number  $C$  defined in the two-triplet model<sup>12</sup> is then

$$\frac{1}{3}C = \tilde{Q} = \tilde{I}_3 + \frac{1}{2}\tilde{Y}. \quad (24)$$

Now it is interesting to note that according to Eq. (22) and Fig. 3, the centers of the three triplets form an antitriplet, equivalent to an antiquark, symmetrically located around the origin. Let us suppose that the nine members of the three triplets  $t_{1\alpha}$ ,  $t_{2\alpha}$ ,  $t_{3\alpha}$ ,  $\alpha = 1, 2, 3$  be combined into a single multiplet  $T = \{t_{i\alpha}\}$ ,  $i = 1, 2, 3$ . We can then imagine two distinct sets of  $SU(3)$  operations on  $T$ . One is the  $SU(3)$  acting on the index  $\alpha$  for each triplet, while the other  $SU(3)$  acts on the index  $i$ , which mixes corresponding members of different triplets.  $T$  is then a representation  $(3,3^*)$  of this group  $\tilde{G} \equiv SU(3)' \times SU(3)''$ .<sup>14</sup> The quantum numbers of  $SU(3)'$  and

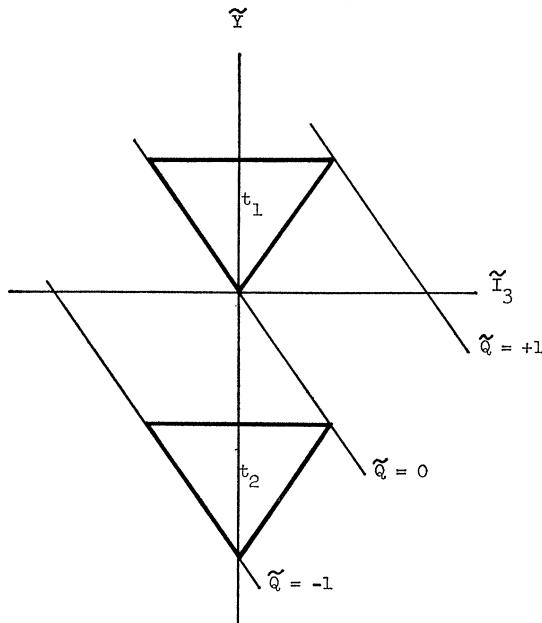


FIG. 2. The two-triplet model.

<sup>14</sup> Such a nonet provides a natural basis for the symmetry of  $SU(9)$ . However, we will not consider it here.

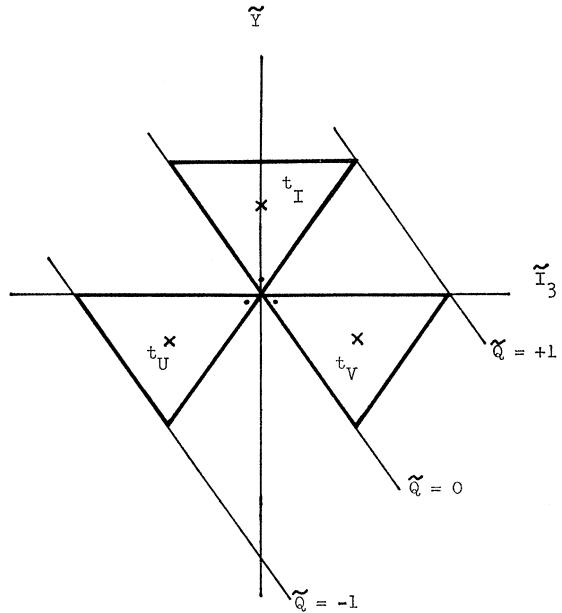


FIG. 3. The three-triplet model.

$SU(3)''$  are identified as  $I_3' = I_3$ ,  $Y' = Y$ ,  $I_3'' = \tilde{I}_3$  and  $Y'' = \tilde{Y}$  in Eq. (22), so that

$$\begin{aligned}\tilde{I}_3 &= I_3' + I_3'', \quad \tilde{Y} = Y' + Y'', \\ \tilde{Q} &= I_3' + I_3'' + \frac{1}{2}Y' + \frac{1}{2}Y'', \\ \frac{1}{3}C &= I_3'' + \frac{1}{2}Y''.\end{aligned}\quad (25)$$

A general representation of  $G$  may be characterized by four numbers  $p'$ ,  $q'$ ,  $p''$ ,  $q''$  so that  $D(p', q', p'', q'') \sim D(p', q') \times D(p'', q'')$ , where  $D(p, q)$  is a representation of  $SU(3)$ . However, in our scheme where the nonet  $T$  is the fundamental field, we do not get all the possible representations of  $G$ . This can be illustrated by means of the triality numbers<sup>5</sup>  $t' = p' - q' \bmod(3)$ ,  $t'' = p'' - q'' \bmod(3)$ . The nonet  $T$  has  $t' = 1$ ,  $t'' = -1$ . All representations constructed out of  $T$  and  $T^*$  then satisfy  $t' = -t''$ .

Let us next consider the meson and baryon states  $\sim TT^*$  and  $\sim TTT$ . The  $SU(3)' \times SU(3)''$  contents of these 81- and 729-plets are

$$\begin{aligned}(3,3^*) \times (3^*, 3) &= (8,1) + (1,1) + (1,8) + (8,8), \\ (3,3^*) \times (3,3^*) \times (3,3^*) &= (1,1) + 2(8,1) + 2(1,8) \\ &\quad + (1,10^*) + (10,1) + 2(8,10^*) + 2(10,8) \\ &\quad + 4(8,8) + (10,10^*).\end{aligned}\quad (26)$$

It is an attractive possibility to postulate at this point that the energy levels are classified according to  $SU(3)''$ . The masses will then depend on the Casimir operators of  $SU(3)''$ . For example, a simple linear form will be

$$m = m_0 + m_2 C_2'' + m_3 C_3'', \quad (27)$$

where  $C_2''$ ,  $C_3''$  are the eigenvalues of quadratic and cubic Casimir operators of  $SU(3)''$ . In particular, we may assume that the main mass splitting comes from  $C_2''$ . Since this increases with the dimensionality of representation, the lowest mass levels will be  $SU(3)''$  singlets. This selects the low-lying meson and baryon states to be (8,1), (1,1) and (8,1), (1,1), (10,1), respectively. In general, all low-lying states will have triality zero,  $t=t''=0$ .

As for the baryon number assignment to the triplets, the simplest possibility would be to assign an equal baryon number, i.e.,  $B=\frac{1}{3}$ , to them. In this case the triplets themselves would be essentially stable, and their nine members would behave like an octet plus a singlet of "heavy baryons" as may be seen from Fig. 3. Another simple possibility may be  $B=\frac{1}{3}+Y''$ , namely  $B=(1,0,0)$  for  $(t_1, t_2, t_3)$ . We expect a mass splitting depending on  $B$  or  $Y''$ , which may be the origin of the Okubo-Gell-Mann mass formula.

The advantage of the three-triplet model is that the  $SU(6)$  symmetry can be easily realized with  $s$ -state triplets. The extended symmetry group becomes now  $SU(6)' \times SU(3)''$ . Since an  $SU(3)''$  singlet is anti-symmetric, the over-all Pauli principle requires the baryon states to be the symmetric  $SU(6)$  56-plet. Other  $SU(6)$  representations such as the 70, will be obtained by bringing in either the orbital angular momentum or the " $\rho$  spin" of the Dirac spinor triplets.

As in the two-triplet model mentioned in the Introduction, the mass formula of the type (27) may be derived dynamically. Instead of the charm number field, we introduce now eight gauge vector fields which behave as (1,8), namely as an octet in  $SU(3)''$ , but as singlets in  $SU(3)'$ . Since their coupling to the individual triplets is proportional to  $\lambda_i''$  [the generators of  $SU(3)''$ ], the interaction energy arising from the exchange of these vector fields will yield the first and second terms of Eq. (27). If these mesons obey again a similar type of mass formula, they will be expected to be massive compared to the ordinary mesons. However, it is not clear whether the resulting short-range character of the interaction can be readily reconciled with the postulated largeness of the interaction energy.

We may characterize the hierarchy of interactions and their symmetries implied by the above model as

follows. First, the *superstrong* interactions responsible for forming baryons and mesons have the symmetry  $SU(3)''$ , and causes large mass splittings between different representations. The scale of mass involved would be comparable or large compared to the baryon mass, namely  $\gtrsim 1$  BeV. The lowest states, i.e.,  $SU(3)''$  singlet states, would split according to  $SU(3)'$ , which would be the  $SU(3)$  group observed among the known baryons and mesons, with their *strong* interactions. The scale of mass splitting would then be  $\lesssim 1$  BeV.

When we go to the massive  $SU(3)''$  nonsinglet states, there may very well be coupling between the two  $SU(3)$  groups similar to the  $L \cdot S$  coupling. The levels should be classified in terms of the three sets of Casimir operators formed out of  $\lambda_i'$ ,  $\lambda_i''$ , and  $\lambda_i=\lambda_i'+\lambda_i''$ , respectively. The splitting due to the coupling would naturally be intermediate between the above two splittings, namely  $\sim 1$  BeV. Because of this coupling, the separate conservation of the two  $SU(3)$  spins,  $I_3'$  and  $Y'$  on the one hand, and  $I_3''$  and  $Y''$  on the other, would be destroyed, and only the sums  $I_3=I_3'+I_3''$  and  $Y=Y'+Y''$  would be conserved under *strong* interactions. This in turn would mean that all the massive states are in general highly unstable, and decay strongly to the low-lying states. (In the two-triplet model, we considered only weak decays of  $C \neq 0$  states. But strong decays are also a possibility as is contemplated here.)

We have discussed here a possible model of baryons and mesons based on three triplets. How can we distinguish this and other different models mentioned already? Certainly different models predict considerably different structure of massive states. These states are characterized by the triality for the quark model, by the charm number for the two-triplet model and by the  $SU(3)''$  representation for the present three-triplet model. If we restrict ourselves to the low-lying states only, however, it seems difficult to distinguish them without making more detailed dynamical assumptions.

#### ACKNOWLEDGMENTS

One of us (M. Y. H.) wishes to thank Professor E. C. G. Sudarshan and Professor A. J. Macfarlane for their encouragement and useful discussions and Professor L. O'Raifeartaigh and J. Kuriyan for helpful comments.



## Self-Consistent Equations Including Exchange and Correlation Effects\*

W. KOHN AND L. J. SHAM

*University of California, San Diego, La Jolla, California*

(Received 21 June 1965)

From a theory of Hohenberg and Kohn, approximation methods for treating an inhomogeneous system of interacting electrons are developed. These methods are exact for systems of slowly varying or high density. For the ground state, they lead to self-consistent equations analogous to the Hartree and Hartree-Fock equations, respectively. In these equations the exchange and correlation portions of the chemical potential of a uniform electron gas appear as additional effective potentials. (The exchange portion of our effective potential differs from that due to Slater by a factor of  $\frac{2}{3}$ .) Electronic systems at finite temperatures and in magnetic fields are also treated by similar methods. An appendix deals with a further correction for systems with short-wavelength density oscillations.

### I. INTRODUCTION

In recent years a great deal of attention has been given to the problem of a homogeneous gas of interacting electrons and its properties have been established with a considerable degree of confidence over a wide range of densities. Of course, such a homogeneous gas represents only a mathematical model, since in all real systems (atoms, molecules, solids, etc.) the electronic density is nonuniform.

It is then a matter of interest to see how properties of the homogeneous gas can be utilized in theoretical studies of inhomogeneous systems. The well-known methods of Thomas-Fermi<sup>1</sup> and the Slater<sup>2</sup> exchange hole are in this spirit. In the present paper we use the formalism of Hohenberg and Kohn<sup>3</sup> to carry this approach further and we obtain a set of self-consistent equations which include, in an approximate way, exchange and correlation effects. They require only a knowledge of the true chemical potential,  $\mu_h(n)$ , of a homogeneous interacting electron gas as a function of the density  $n$ .

We derive two alternative sets of equations [Eqs. (2.8) and (2.22)] which are analogous, respectively, to the conventional Hartree and Hartree-Fock equations, and, although they also include correlation effects, they are no more difficult to solve.

The local effective potentials in these equations are unique in a sense which is described in Sec. II. In particular, we find that the Slater exchange-hole potential, besides its omission of correlation effects, is too large by a factor of  $\frac{3}{2}$ .

Apart from work on the correlation energy of the homogeneous electron gas, most theoretical many-body studies have been concerned with elementary excitations and as a result there has been little recent progress in the theory of cohesive energies, elastic constants, etc., of real (i.e., inhomogeneous) metals and alloys. The methods proposed here offer the hope of new progress in this latter area.

\* Supported in part by the U. S. Office of Naval Research.

<sup>1</sup> L. H. Thomas, Proc. Cambridge Phil. Soc. 23, 542 (1927); E. Fermi, Z. Physik 48, 73 (1928).

<sup>2</sup> J. C. Slater, Phys. Rev. 81, 385 (1951).

<sup>3</sup> P. Hohenberg and W. Kohn, Phys. Rev. 136, B864 (1964); referred to hereafter as HK.

In Secs. III and IV, we describe the necessary modifications to deal with the finite-temperature properties and with the spin paramagnetism of an inhomogeneous electron gas.

Of course, the simple methods which are here proposed in general involve errors. These are of two general origins<sup>4</sup>: a too rapid variation of density and, for finite systems, boundary effects. Refinements aimed at reducing the first type of error are briefly discussed in Appendix II.

### II. THE GROUND STATE

#### A. Local Effective Potential

It has been shown<sup>5</sup> that the ground-state energy of an interacting inhomogeneous electron gas in a static potential  $v(r)$  can be written in the form

$$E = \int v(r)n(r) dr + \frac{1}{2} \iint \frac{n(r)n(r')}{|r-r'|} dr dr' + G[n], \quad (2.1)$$

where  $n(r)$  is the density and  $G[n]$  is a universal functional of the density. This expression, furthermore, is a minimum for the correct density function  $n(r)$ . In this section we propose first an approximation for  $G[n]$ , which leads to a scheme analogous to Hartree's method but contains the major part of the effects of exchange and correlation.

We first write

$$G[n] \equiv T_s[n] + E_{xc}[n], \quad (2.2)$$

where  $T_s[n]$  is the kinetic energy of a system of non-interacting electrons with density<sup>6</sup>  $n(r)$  and  $E_{xc}[n]$  is, by our definition, the exchange and correlation energy of an interacting system with density  $n(r)$ . For an arbitrary  $n(r)$ , of course, one can give no simple exact expression for  $E_{xc}[n]$ . However, if  $n(r)$  is sufficiently slowly varying, one can show<sup>3</sup> that

$$E_{xc}[n] = \int n(r)\epsilon_{xc}(n(r)) dr, \quad (2.3)$$

<sup>4</sup> W. Kohn and L. J. Sham, Phys. Rev. 137, A1697 (1965).

<sup>5</sup> For such a system it follows from HK that the kinetic energy is in fact a unique functional of the density.

where  $\epsilon_{xc}(n)$  is the exchange and correlation energy per electron of a uniform electron gas of density  $n$ . Our sole approximation consists of assuming that (2.3) constitutes an adequate representation of exchange and correlation effects in the systems under consideration. We shall regard  $\epsilon_{xc}$  as known from theories of the homogeneous electron gas.<sup>6</sup>

From the stationary property of Eq. (2.1) we now obtain, subject to the condition

$$\int \delta n(\mathbf{r}) d\mathbf{r} = 0, \quad (2.4)$$

the equation

$$\int \delta n(\mathbf{r}) \left\{ \varphi(\mathbf{r}) + \frac{\delta T_s[n]}{\delta n(\mathbf{r})} + \mu_{xc}(n(\mathbf{r})) \right\} d\mathbf{r} = 0; \quad (2.5)$$

here

$$\varphi(\mathbf{r}) = v(\mathbf{r}) + \int \frac{n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}', \quad (2.6)$$

and

$$\mu_{xc}(n) = d(n\epsilon_{xc}(n))/dn \quad (2.7)$$

is the exchange and correlation contribution to the chemical potential of a uniform gas of density  $n$ .

Equations (2.4) and (2.5) are precisely the same as one obtains from the theory of Ref. 3 when applied to a system of noninteracting electrons, moving in the given potential  $\varphi(\mathbf{r}) + \mu_{xc}(n(\mathbf{r}))$ . Therefore, for given  $\varphi$  and  $\mu$ , one obtains the  $n(\mathbf{r})$  which satisfies these equations simply by solving the one-particle Schrödinger equation

$$\{-\frac{1}{2}\nabla^2 + [\varphi(\mathbf{r}) + \mu_{xc}(n(\mathbf{r}))]\}\psi_i(\mathbf{r}) = \epsilon_i \psi_i(\mathbf{r}), \quad (2.8)$$

and setting

$$n(\mathbf{r}) = \sum_{i=1}^N |\psi_i(\mathbf{r})|^2, \quad (2.9)$$

where  $N$  is the number of electrons.

It is physically very satisfactory that  $\mu_{xc}$  appears in Eq. (2.8) as an additional effective potential so that gradients of  $\mu_{xc}$  lead to forces on the electron fluid in a manner familiar from thermodynamics.

Equations (2.6)–(2.9) have to be solved self-consistently: One begins with an assumed  $n(\mathbf{r})$ , constructs  $\varphi(\mathbf{r})$  from (2.6) and  $\mu_{xc}$  from (2.7), and finds a new  $n(\mathbf{r})$  from (2.8) and (2.9). The energy is given by

$$E = \sum_i^N \epsilon_i - \frac{1}{2} \int \int \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r} d\mathbf{r}' + \int n(\mathbf{r}) [\epsilon_{xc}(n(\mathbf{r})) - \mu_{xc}(n(\mathbf{r}))] d\mathbf{r}. \quad (2.10)$$

The results of our procedure are exact in two limiting cases:

(a) *Slowly varying density.* This regime is characterized by the condition  $r_s/r_0 \ll 1$ , where  $r_s$  is the Wigner-

<sup>6</sup> For a review see D. Pines, *Elementary Excitations in Solids* (W. A. Benjamin, Inc., New York, 1963).

Seitz radius and  $r_0$  is a typical length over which there is an appreciable change in density. In this case, as shown in HK, we can expand the true exchange and correlation energy as follows:

$$E_{xc}[n] = \int \epsilon_{xc}(n) n d\mathbf{r} + \int \epsilon_{xc}^{(2)}(n) |\nabla n|^2 d\mathbf{r} + \dots, \quad (2.11)$$

where  $\epsilon_{xc}^{(2)}$  is the exchange and correlation portion of the second term in the energy expansion in powers of the gradient operator. In this regime we may similarly expand  $T_s[n]$  in the form

$$T_s[n] = \int \frac{3}{10} (3\pi^2 n)^{2/3} n d\mathbf{r} + \int t^{(2)}(n) |\nabla n|^2 d\mathbf{r} + \dots. \quad (2.12)$$

From HK, especially Sec. III 2, we have the following expression for the energy:

$$E_v[n] = \int v(\mathbf{r}) n(\mathbf{r}) d\mathbf{r} + \frac{1}{2} \int \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r} d\mathbf{r}' + \int g_0(n) d\mathbf{r} + \int g_2^{(2)}(n) |\nabla n|^2 d\mathbf{r} + \dots, \quad (2.13)$$

where

$$g_0(n) = \{\frac{3}{10} (3\pi^2 n)^{2/3} + \epsilon_{xc}(n)\} n, \quad (2.14)$$

and

$$g_2^{(2)}(n) = \{\epsilon_{xc}^{(2)}(n) + t^{(2)}(n)\} n. \quad (2.15)$$

Since in our approximation (2.3), the  $|\nabla|^2$  term of Eq. (2.11) is neglected, it is clear that for a gas of slowly varying density our expression (2.10) for the energy has errors of the order  $|\nabla|^2$ , or equivalently, of the order  $r_0^{-2}$ .

Surprisingly, our procedure determines the density with greater accuracy, the errors being of order  $|\nabla|^4$ . This is shown in Appendix I.

At this point a comparison of our procedure and that of Slater<sup>2</sup> may be appropriate. For one thing, Slater's original work does not include correlation effects.<sup>7</sup> But even the exchange correction is different from ours. To obtain Slater's exchange correction, one may begin by writing the Hartree-Fock exchange operator in the form of an equivalent potential acting on the  $k$ th wave function

$$v_{xk}(\mathbf{r}) = - \sum_{k'=1}^N \int \frac{\psi_k^*(\mathbf{r}) \psi_{k'}^*(\mathbf{r}') \psi_{k'}(\mathbf{r}) \psi_k(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}' / \psi_k^*(\mathbf{r}) \psi_k(\mathbf{r}), \quad (2.16)$$

<sup>7</sup> Subsequent to the original paper by Slater, there have been several attempts to add correlation corrections: S. Olszewski, Phys. Rev. 121, 42 (1961); J. E. Robinson, F. Bassani, B. S. Knox, and J. R. Schrieffer, Phys. Rev. Letters 9, 215 (1962); W. A. Harrison, Phys. Rev. 136, A1107 (1964); S. Lundqvist and C. W. Ufford, Phys. Rev. 139, A1 (1965).

where the symbols  $\mathbf{r}$  and  $\mathbf{r}'$  are understood to include electron spin coordinates and integration is understood to include summation over spin coordinates. One next assumes that the wave functions can be approximated by plane waves which results in

$$v_{xk}(\mathbf{r}) = -\frac{k_F(\mathbf{r})}{\pi} \left[ 1 + \frac{k_F^2(\mathbf{r}) - k^2}{2kk_F(\mathbf{r})} \ln \left| \frac{k+k_F(\mathbf{r})}{k-k_F(\mathbf{r})} \right| \right], \quad (2.17)$$

where  $k_F(\mathbf{r}) \equiv \{3\pi^2n(\mathbf{r})\}^{1/3}$ . Finally, one averages  $v_{xk}$  over the occupied state  $k$ , which results in

$$v_x(\mathbf{r}) = -(3/2\pi)\{3\pi^2n(\mathbf{r})\}^{1/3}. \quad (2.18)$$

In our procedure (neglecting correlation) we obtain, in place of Slater's  $v_x$

$$\mu_x(\mathbf{r}) = -(1/\pi)\{3\pi^2n(\mathbf{r})\}^{1/3}, \quad (2.19)$$

smaller by a factor of  $\frac{2}{3}$ . From the discussion in Appendix I, it follows that while  $\mu_x$  gives the exchange correction of the density correct to order  $|\nabla|^2$ , inclusive,  $v_x$  [as indeed any other function of  $n(\mathbf{r})$ ] leads to errors of order  $|\nabla|^2$ . The same comment applies to any extension of Slater's exchange to include correlation in the self-consistent potential.

We may note that our result is equivalent to taking, not the average of (2.17), but rather its value at  $k=k_F(\mathbf{r})$ ; i.e., the effective exchange potential for a state at the top of the Fermi distributions. This is physically understandable since density adjustments come about by redistribution of the electrons near the Fermi level.

(b) *High density.* This regime is characterized by the condition  $r_s/a_0 \ll 1$ , where  $a_0$  is the Bohr radius. In this case, the entire exchange and correlation energy is smaller than the kinetic energy by a factor of order  $(r_s/a_0)$  and hence our inaccuracy in representing these portions becomes negligible.

The reader will have noticed that while in Eq. (2.3) we approximate the exchange and correlation energy by the expression valid for a slowly varying density, we made no approximation for the kinetic-energy functional  $T_s[n]$  of Eq. (2.2). This procedure is responsible for the exactness of the high-density limit, even when the density is rapidly varying, such as in the vicinity of an atomic nucleus.

We now make a few further remarks about our approximation. If in Eq. (2.2), we had approximated  $T_s[n]$  by its form appropriate to a system of slowly varying density,

$$T_s[n] \rightarrow \int \frac{3}{16}(3\pi^2n)^{2/3}n d\mathbf{r}, \quad (2.20)$$

we would have been led to the generalization of the Thomas-Fermi method suggested by Lewis.<sup>8</sup> This method shares with the Thomas-Fermi method two shortcomings: (1) It leads to an infinite density near

an atomic nucleus, and (2) it does not lead to quantum density oscillations,<sup>4</sup> such as the density fluctuations due to atomic shell structures. By not making the replacement (2.20), we avoid both of these shortcomings.

Let us now qualitatively discuss the appropriateness of our procedure for various classes of electronic systems.

In atoms and molecules one can distinguish three regions: (1) A region near the atomic nucleus, where the electronic density is high and therefore, in view of case (b) above, we expect our procedure to be satisfactory. (2) The main "body" of the charge distribution where the electronic density  $n(\mathbf{r})$  is relatively slowly varying, so that our approximation (2.3) for  $\epsilon_{xc}$  is expected to be satisfactory as discussed in case (a) above. (3) The "surface" of atoms and the overlap regions in molecules. Here our approximation (2.3) has no validity and therefore we expect this region to be the main source of error. We do not expect an accurate description of chemical binding. In large atoms, of course, this "surface" region becomes of less importance. (The surface is more satisfactorily handled in the nonlocal method described under B below.)

For metals, alloys, and small-gap insulators we have, of course, no surface problem and we expect our approximation (2.3) to give a good representation of exchange and correlation effects. In large-gap insulators, however, the actual correlation energy will be considerably reduced compared to that of a homogeneous electron gas of the same density.

## B. Nonlocal Effective Potential

Instead of the Hartree-type procedure discussed in Sec. IIA it is also possible to obtain a scheme which includes exchange effects exactly. We write in place of Eq. (2.3)

$$E_{xc}[n] = E_x[n] + \int n(\mathbf{r})\epsilon_c(n(\mathbf{r})) d\mathbf{r} \quad (2.21)$$

where  $E_x[n]$  is the exchange energy of a Hartree-Fock system of density  $n(\mathbf{r})$  and  $\epsilon_c(n)$  is the correlation energy per particle of a homogeneous electron gas. Applying this ansatz in conjunction with Eq. (2.2) and the stationary property of (2.1) leads to the following system of equations:

$$\{-\frac{1}{2}\nabla^2 + \varphi(\mathbf{r}) + \mu_c(\mathbf{r})\}\psi_i(\mathbf{r}) - \int \frac{n_1(\mathbf{r}, \mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} \psi_i(\mathbf{r}') d\mathbf{r}' = \epsilon_i \psi_i(\mathbf{r}), \quad (2.22)$$

where

$$\mu_c = d(n\epsilon_c)/dn, \quad (2.23)$$

$$n_1(\mathbf{r}, \mathbf{r}') = \sum_{j=1}^N \psi_j(\mathbf{r})\psi_j^*(\mathbf{r}'), \quad (2.24)$$

and  $\varphi(\mathbf{r})$ ,  $n(\mathbf{r})$  are defined as before, Eqs. (2.6) and (2.9).

<sup>8</sup> H. W. Lewis, Phys. Rev. 111, 1554 (1958).

The energy is now

$$\begin{aligned} E = & \sum_i^N \epsilon_i - \frac{1}{2} \int \int \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r} d\mathbf{r}' \\ & + \frac{1}{2} \int \int \frac{n_1(\mathbf{r}, \mathbf{r}')n_1(\mathbf{r}', \mathbf{r})}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r} d\mathbf{r}' \\ & + \int n(\mathbf{r}) \{ \epsilon_e(n(\mathbf{r})) - \mu_e(n(\mathbf{r})) \} d\mathbf{r}. \end{aligned} \quad (2.25)$$

This procedure may be regarded as a Hartree-Fock method corrected for correlation effects. It is no more complicated than the uncorrected Hartree-Fock method but, because of the nonlocal operator appearing in Eq. (2.22), very much more complicated than the method described in Sec. IIA. Since at least exchange effects are now treated exactly we must expect, in general, more accurate results than from the method of Sec. IIA. In particular, near the surface of an atom the effective potential now is correctly ( $-1/r$ ) whereas in Sec. IIA it approaches zero much faster. Even here, however, correlation effects are not correctly described near the surface.

### III. FREE ENERGY; SPECIFIC HEAT

We can generalize the consideration of the ground state to finite temperature ensembles by using the finite temperature generalization of Eq. (2.1) given by Mermin.<sup>9</sup> He has shown that the grand canonical potential can be written in the form

$$\Omega = \int v(\mathbf{r})n(\mathbf{r}) d\mathbf{r} + \frac{1}{2} \int \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r} d\mathbf{r}' + G[n] - \mu \int n(\mathbf{r}) d\mathbf{r}, \quad (3.1)$$

where  $G[n]$  is a unique functional of the density at a given temperature  $\tau$  and  $\mu$  is the chemical potential. For the correct  $n$  this quantity is a minimum.

In analogy with (2.2) we now write

$$G[n] = G_s[n] + F_{xc}[n]; \quad (3.2)$$

here

$$G_s[n] \equiv T_s[n] - \tau S_s[n], \quad (3.3)$$

where  $T_s[n]$  and  $S_s[n]$  are, respectively, the kinetic energy and entropy of noninteracting electrons with density  $n(\mathbf{r})$  at a temperature  $\tau$ ; and  $F_{xc}[n]$  is, by definition, the exchange and correlation contribution to the free energy. For the latter quantity, we make the approximation

$$F_{xc}[n] = \int n(\mathbf{r})f_{xc}(n(\mathbf{r})) d\mathbf{r}, \quad (3.4)$$

where  $f_{xc}(n)$  is the exchange and correlation contribution to the free energy per electron of a uniform electron

gas of density  $n$ ; i.e.,

$$f_{xc}(n) \equiv f(n) - f_0(n), \quad (3.5)$$

where  $f$  and  $f_0$  are the free energies per electron of an interacting and noninteracting gas, respectively.

$$0 = \varphi(\mathbf{r}) + (\delta G_s[n]/\delta n(\mathbf{r})) + \mu_{xc}(n(\mathbf{r})) - \mu, \quad (3.6)$$

where  $\varphi(\mathbf{r})$  is given, as before, by Eq. (2.6) and

$$\mu_{xc}(n) \equiv d(n f_{xc}(n))/dn. \quad (3.7)$$

Equation (3.6) is identical to the corresponding equation for a system of noninteracting electrons in the effective potential  $\varphi + \mu_{xc}$ . Its solution is therefore determined by the following system of equations:

$$\{-\frac{1}{2}\nabla^2 + \varphi(\mathbf{r}) + \mu_{xc}(n(\mathbf{r}))\}\psi_i = \epsilon_i \psi_i, \quad (3.8)$$

and

$$n(\mathbf{r}) = \sum_{i=1}^N |\psi_i(\mathbf{r})|^2 / \{e^{(\epsilon_i - \mu)/k\tau} + 1\}. \quad (3.9)$$

$\mu$  is determined as usual by the total number of particles from Eq. (3.9). This value also represents our approximation for the chemical potential of the interacting system.

Of special interest for metals and alloys is the low-temperature heat capacity. This may be obtained by making an expansion, in powers of  $\tau$ , of the above system of equations. An equivalent, but more convenient, method is as follows: From thermodynamics and Eq. (3.1) we have

$$\begin{aligned} S[n] \equiv & -\frac{\partial}{\partial \tau} (\Omega + \mu N)_V = -\int \left\{ \varphi(\mathbf{r}) + \frac{\delta G}{\delta n(\mathbf{r})} \right\} \\ & \times \left( \frac{\partial n(\mathbf{r})}{\partial \tau} \right)_V d\mathbf{r} - \left( \frac{\partial G[n]}{\partial \tau} \right)_{n(\mathbf{r}), V}. \end{aligned} \quad (3.10)$$

The integral vanishes because of the stationary property of  $\Omega$ , so that

$$S[n] = -(\delta G[n]/\delta \tau)_{n(\mathbf{r}), V}. \quad (3.11)$$

The same argument, applied to a system of noninteracting electrons of density  $n(\mathbf{r})$ , gives

$$S_s[n] = -(\delta G_s[n]/\delta \tau)_{n(\mathbf{r}), V}. \quad (3.12)$$

Combining Eqs. (3.11), (3.12), (3.2), and (3.4), we obtain

$$S[n] = S_s[n] + \int n(\mathbf{r}) (\partial f_{xc}(n)/\partial \tau)_{n(\mathbf{r}), V} d\mathbf{r}. \quad (3.13)$$

For small  $\tau$  it is well known that  $S_s$  is given by

$$S_s[n] = N \frac{1}{3} \pi^2 k^2 \tau g_s(\mu), \quad (3.14)$$

where  $g_s$  is the single-particle density of states in the effective potential  $\varphi + \mu_{xc}$  at zero temperature; further,

$$(\partial f_{xc}(n)/\partial \tau)_{n(\mathbf{r}), V} = \frac{1}{3} \pi^2 k^2 \tau [g(\mu_h(n)) - g_0(\mu_0(n))], \quad (3.15)$$

<sup>9</sup> N. D. Mermin, Phys. Rev. 137, A1441 (1965).

where  $\mu_h(n)$  and  $\mu_0(n)$  are, respectively, the chemical potentials of an interacting and a noninteracting homogeneous gas of density  $n$ , and  $g$  and  $g_0$  are the respective densities of states.<sup>10</sup>

It follows immediately that the low-temperature heat capacity is given by

$$C_v = \gamma\tau, \quad (3.16)$$

where

$$\gamma = \frac{1}{3}\pi^2 k^2 \left[ N g_s(\mu) + \int n(\mathbf{r}) \{g(\mu_h(n)) - g_0(\mu_0(n))\} d\mathbf{r} \right]. \quad (3.17)$$

We shall not present a treatment, analogous to Sec. II B, in which exchange effects are included exactly. The development is straightforward but leads to a well-known divergence in the low-temperature specific heat.

#### IV. SPIN SUSCEPTIBILITY

To obtain a theory of the spin susceptibility of an electron gas, we first extend the theory of HK to include the effects of spin interaction with an external magnetic field. The result is that if we take the field in the  $z$  direction and write the magnetic-moment density as

$$m(\mathbf{r}) = -(1/2c)\langle 0 | \psi_{\uparrow}^*(\mathbf{r})\psi_{\uparrow}(\mathbf{r}) - \psi_{\downarrow}^*(\mathbf{r})\psi_{\downarrow}(\mathbf{r}) | 0 \rangle, \quad (4.1)$$

the ground-state energy can be written in the form

$$E_{v,H} = \int \{v(\mathbf{r})n(\mathbf{r}) - H(\mathbf{r})m(\mathbf{r})\} d\mathbf{r} + \frac{1}{2} \int \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r} d\mathbf{r}' + G[n(\mathbf{r}), m(\mathbf{r})], \quad (4.2)$$

where  $G$  is a universal functional of  $n$  and  $m$ , and the correct  $m(\mathbf{r})$ ,  $n(\mathbf{r})$  make (4.2) a minimum.

For small  $m$  we expand  $G$  in the form

$$G = G[n] + \frac{1}{2} \int G(\mathbf{r}, \mathbf{r}'; [n])m(\mathbf{r})m(\mathbf{r}') d\mathbf{r} d\mathbf{r}' + \dots; \quad (4.3)$$

the linear term vanishes for a paramagnetic system in which  $m \equiv 0$  when  $H \equiv 0$ . From the stationary property of (4.2) we find, for small  $H$ , that  $n$  is unchanged to first order and that

$$-H(\mathbf{r}) + \int G(\mathbf{r}, \mathbf{r}'; [n])m(\mathbf{r}') d\mathbf{r}' = 0, \quad (4.4)$$

where  $n$  is the zero-field density. We now formally invert this equation, which gives

$$m(\mathbf{r}) = \int G^{-1}(\mathbf{r}, \mathbf{r}'; [n])H(\mathbf{r}') d\mathbf{r}'. \quad (4.5)$$

For a uniform field this gives for the susceptibility

$$\chi[n] = \frac{1}{V} \frac{\partial}{\partial H} \int m(\mathbf{r}) d\mathbf{r} = \int G^{-1}(\mathbf{r}, \mathbf{r}'; [n]) d\mathbf{r} d\mathbf{r}'. \quad (4.6)$$

<sup>10</sup> J. M. Luttinger, Phys. Rev. 119, 1153 (1960).

So far everything is formal and exact. We now write, in the spirit of the previous sections,

$$G^{-1}(\mathbf{r}, \mathbf{r}'; [n]) \equiv G_s^{-1}(\mathbf{r}, \mathbf{r}'; [n]) + G_{xc}^{-1}(\mathbf{r}, \mathbf{r}'; [n]). \quad (4.7)$$

The second term we approximate as for a slowly varying gas, which gives

$$\chi[n] = \chi_s[n] + \frac{1}{V} \int [\chi(n(\mathbf{r})) - \chi_0(n(\mathbf{r}))] d\mathbf{r}, \quad (4.8)$$

where

$$\chi_s[n] = (1/2c)^2 (N/V) \times g_s(\mu), \quad (4.9)$$

and  $\chi(n)$ ,  $\chi_0(n)$  are, respectively, the susceptibilities for uniform systems with and without interactions.

#### APPENDIX I: GRADIENT EXPANSION OF THE DENSITY

In this Appendix we show that for a system of slowly varying density our procedure gives the density correct to order  $|\nabla|^2$  inclusive. When dealing with such a system we may proceed in two entirely equivalent ways: (1) We can solve the self-consistent equations, Eqs. (2.8) and (2.9), for  $n(\mathbf{r})$ , and (2) we can go back to the underlying variational principle (2.5), make a gradient expansion and determine  $n(\mathbf{r})$  directly. We shall here follow the second route to estimate the errors in  $n(\mathbf{r})$ .

From (2.5) and the expansion (2.12) of  $T_s[n]$ , we obtain

$$\mu = \varphi(\mathbf{r}) + \mu_h(n) - t^{(2)''}(n) |\nabla n|^2 - 2t^{(2)}(n) \nabla^2 n + O(\nabla^4), \quad (A1.1)$$

where  $\mu$  is the chemical potential [cf. HK, Eq. (68)]. Note however that because of our approximation of keeping only the first term in (2.11), some other contributions of order  $|\nabla|^2$  are missing in (A1.1).

To solve (A1.1), let us write the external charge density as

$$n_{\text{ext}}(\mathbf{r}) \equiv f_0(\mathbf{r}/r_0), \quad (A1.2)$$

where  $r_0 \rightarrow \infty$  (slow spatial variation), and try the ansatz

$$n(\mathbf{r}) = n_0(\mathbf{r}) + n_1(\mathbf{r}), \quad (A1.3)$$

where

$$n_0(\mathbf{r}) = f_0(\mathbf{r}/r_0) \quad (A1.4)$$

exactly neutralizes the external charge and  $n_1$  is assumed to approach zero as  $r_0 \rightarrow \infty$ . Neglecting, for the moment, the terms of order  $|\nabla|^2$  in (A1.1) and substituting (A1.3) into (A1.1), we obtain

$$\mu = \int \frac{n_1(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}' + \mu_h(n_0) + n_1(\mathbf{r})\mu_h'(n_0) + O(n_1^2). \quad (A1.5)$$

Now define

$$\mathbf{R} = \mathbf{r}/r_0, \quad (A1.6)$$

and write

$$n_1(\mathbf{r}) \equiv f_1(\mathbf{R}). \quad (A1.7)$$

With this notation, (A1.5) becomes

$$\mu = r_0^2 \int \frac{f_1(\mathbf{R}')}{|\mathbf{R} - \mathbf{R}'|} d\mathbf{R}' + \mu_h(f_0(\mathbf{R})) + f_1(\mathbf{R})\mu'_h(f_0(\mathbf{R}')) + O(f_1^2). \quad (\text{A1.8})$$

We may now write

$$f_1(\mathbf{R}) = (1/r_0^2)f_1^{(2)}(\mathbf{R}) + (1/r_0^4)f_1^{(4)}(\mathbf{R}) + \dots, \quad (\text{A1.9})$$

and

$$\mu = \mu^{(0)} + (1/r_0^2)\mu^{(2)} + \dots \quad (\text{A1.10})$$

The first term of Eq. (A1.9) is correctly determined by Eq. (A1.8) and not affected either by the inclusion of terms of order  $\nabla^2$  in (A1.5) or by the terms of order  $f_1^2$  in (A1.8). Hence, in spite of the errors of order  $\nabla^2$  in (A1.1), the density given by our procedure is correct to order  $1/r_0^2$  or  $|\nabla|^2$ , inclusive. Equation (A1.8) shows that this curious result stems from the infinite range of the Coulomb interaction.

## APPENDIX II: EFFECT OF RAPID DENSITY OSCILLATION ON EXCHANGE AND CORRELATION

In Eq. (2.3), we approximated  $E_{xc}[n]$  by the first term in the gradient expansion. In actual physical systems, there are quantum density oscillations<sup>4</sup> whose effects on exchange and correlation are not included in the approximation (2.3). Now we put forward a correction to (2.3) to include such effects.

In HK, the gradient expression for the energy functional is partially summed such that it is also correct for a system of almost constant density<sup>1</sup> even when the density fluctuations are of short wavelength<sup>11</sup>:

$$G[n] = \int g_0(n(\mathbf{r})) d\mathbf{r} - \frac{1}{2} \int K(\mathbf{r} - \mathbf{r}'; n(\bar{\mathbf{r}})) \times \{n(\mathbf{r}) - n(\mathbf{r}')\}^2 d\mathbf{r} d\mathbf{r}', \quad (\text{A2.1})$$

where  $K(\mathbf{r} - \mathbf{r}'; n)$  is determined by the polarizability of a homogeneous electron gas at density  $n$ , and  $\bar{\mathbf{r}} = \frac{1}{2}(\mathbf{r} + \mathbf{r}')$ . To the same approximation,

$$E_{xc}[n] = \int n(\mathbf{r}) \epsilon_{xc}(n(\mathbf{r})) d\mathbf{r} - \frac{1}{2} \int K_{xc}(\mathbf{r} - \mathbf{r}'; n(\bar{\mathbf{r}})) \times \{n(\mathbf{r}) - n(\mathbf{r}')\}^2 d\mathbf{r} d\mathbf{r}' \quad (\text{A2.2})$$

where  $K_{xc}$  is the difference between  $K$  of the interacting homogeneous gas and that of the noninteracting gas at the same density. We believe that for an infinite system,

<sup>11</sup> The second term of HK, Eq. (83) is in error; it should be

$$-\frac{1}{2} \int K(\mathbf{r}'; n(\mathbf{r})) \{n(\mathbf{r} + \frac{1}{2}\mathbf{r}') - n(\mathbf{r} - \frac{1}{2}\mathbf{r}')\}^2 d\mathbf{r}'.$$

The kernel  $K$  has the same meaning as in HK.

such as a metal or an alloy, the second term on the right-hand side of (A2.2) accounts adequately for the effect of rapid density change on exchange and correlation.

This  $E_{xc}[n]$  again leads to a set of Hartree-type equations like Eq. (2.8), with an addition to the effective potential given by

$$\begin{aligned} & -\frac{1}{2} \int \{\partial K_{xc}(\mathbf{r}'; n(\mathbf{r}))/\partial n(\mathbf{r})\} \\ & \times \{n(\mathbf{r} + \frac{1}{2}\mathbf{r}') - n(\mathbf{r} - \frac{1}{2}\mathbf{r}')\}^2 d\mathbf{r}' \\ & - 2 \int K_{xc}(\mathbf{r} - \mathbf{r}'; n(\bar{\mathbf{r}})) \{n(\mathbf{r}) - n(\mathbf{r}')\} d\mathbf{r}'. \end{aligned} \quad (\text{A2.3})$$

Note that in the random-phase approximation  $K_{xc}$  vanishes. Hence, in a calculation which includes the effective potential (A2.3), we need reliable estimates of  $K_{xc}$ , calculated beyond the random-phase approximation, which are not available at present.

The addition of (A2.3) to the effective potential obviously makes the solution of the self-consistent equations much more difficult. However, assuming that the modification of  $n(\mathbf{r})$  produced by this term is small, one may calculate  $n(\mathbf{r})$  and  $E$  first without including it, and then, because of the stationary property, Eq. (2.5), one can obtain the correction to the energy by evaluating the second term in (A2.2) with the unmodified density.

*Note added in proof.* We should like to point out that it is possible, formally, to replace the many-electron problem by an *exactly* equivalent set of self-consistent one-electron equations. This is accomplished quite simply by using the expression (2.2) [without the approximation (2.3)] in the energy variational principle. This leads to a set of equations, analogous to Eqs. (2.4)–(2.9), but with  $\mu_{xc}(n)$  replaced by an effective one-particle potential  $v_{xc}$ , defined formally as

$$v_{xc}(\mathbf{r}) \equiv \delta E_{xc}[n]/\delta n(\mathbf{r}).$$

Of course, an explicit form of  $v_{xc}$  can be obtained only if the functional  $E_{xc}[n]$ , which includes all many-body effects, is known. This effective potential will reproduce the exact density and the exact total energy is then given by

$$\begin{aligned} E = & \sum_i^N \epsilon_i - \frac{1}{2} \int \int \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r} d\mathbf{r}' + E_{xc}[n] \\ & - \int v_{xc}(\mathbf{r}) n(\mathbf{r}) d\mathbf{r}. \end{aligned}$$

Of course, if we make the approximation (2.3) for  $E_{xc}$  the above exact formulation reverts to the approximate theory of Sec. II.

# Quantum Theory of Gravity. I. The Canonical Theory\*

BRYCE S. DEWITT

*Institute for Advanced Study, Princeton, New Jersey*

and

*Department of Physics, University of North Carolina, Chapel Hill, North Carolina†*

(Received 25 July 1966; revised manuscript received 9 January 1967)

Following an historical introduction, the conventional canonical formulation of general relativity theory is presented. The canonical Lagrangian is expressed in terms of the extrinsic and intrinsic curvatures of the hypersurface  $x^0 = \text{constant}$ , and its relation to the asymptotic field energy in an infinite world is noted. The distinction between finite and infinite worlds is emphasized. In the quantum theory the primary and secondary constraints become conditions on the state vector, and in the case of finite worlds these conditions alone govern the dynamics. A resolution of the factor-ordering problem is proposed, and the consistency of the constraints is demonstrated. A 6-dimensional hyperbolic Riemannian manifold is introduced which takes for its metric the coefficient of the momenta in the Hamiltonian constraint. The geodesic incompleteness of this manifold, owing to the existence of a frontier of infinite curvature, is demonstrated. The possibility is explored of relating this manifold to an infinite-dimensional manifold of 3-geometries, and of relating the structure of the latter manifold in turn to the dynamical behavior of space-time. The problem is approached through the WKB approximation and Hamilton-Jacobi theory. Einstein's equations are revealed as geodesic equations in the manifold of 3-geometries, modified by the presence of a "force term." The classical phenomenon of gravitational collapse shows that the force term is not powerful enough to prevent the trajectory of space-time from running into the frontier. The as-yet unresolved problem of determining when the collapse phenomenon represents a real barrier to the quantum-state functional is briefly discussed, and a boundary condition at the barrier is proposed. The state functional of a finite world can depend only on the 3-geometry of the hypersurface  $x^0 = \text{constant}$ . The label  $x^0$  itself is irrelevant, and "time" must be determined intrinsically. A natural definition for the inner product of two such state functionals is introduced which, however, encounters difficulties with negative probabilities owing to the barrier boundary condition. In order to resolve these difficulties, a simplified model, the quantized Friedmann universe, is studied in detail. In order to obtain nonstatic wave functions which resemble a universe evolving, it is necessary to introduce a clock. In order that the combined wave functions of universe-cum-clock be normalizable, it turns out that the periods of universe and clock must be commensurable. Wave packets exhibiting quasiclassical behavior are constructed, and attention is called to the phenomenological character of "time." The inner-product definition is rescued from its negative-probability difficulties by making use of the fact that probability flows in a closed finite circuit in configuration space. The article ends with some speculations on the uniqueness of the state functional of the actual universe. It is suggested that a viewpoint due to Everett should be adopted in its interpretation.

## 1. INTRODUCTION

**A**LMOST as soon as quantum field theory was invented by Heisenberg, Pauli, Fock, Dirac, and Jordan, attempts were made to apply it to fields other than the electromagnetic field which had given it—and indeed quantum mechanics itself—birth. In 1930 Rosenfeld<sup>1</sup> applied it to the gravitational field which, at the time, was still regarded as the *other* great entity of Nature. Rosenfeld was the first to note some of the special technical difficulties involved in quantizing gravity and made some early attempts to develop general methods for handling them. As an application of his methods he computed the gravitational self-energy of a photon in lowest order of perturbation theory. He obtained a quadratically divergent result, confirming that the divergence malady of field theory, which had already been discovered in connection with the electron's electromagnetic self-energy, was widespread and deep seated. It is tempting, and perhaps no longer pre-

mature, to read into Rosenfeld's result a forecast that quantum gravodynamics was destined, from the very beginning, to be inextricably linked with the difficult issues lying at the theoretical foundations of particle physics.

During physics's great boom of the thirties the difficult issues of field theory were inevitably often bypassed. Moreover, it was recognized early that as far as the gravitational field is concerned its quanta (assuming they exist) can produce no *observable* effects until energies of the order of  $10^{28}$  eV are reached, this fantastic energy corresponding to the so-called "Planck length"  $(\hbar G/c^3)^{1/2} \approx 10^{-33}$  cm, where  $G$  is the gravitation constant. Hence, after Rosenfeld's initial studies years passed before anything essentially new was done in quantum gravodynamics, and even today interest in this area of research is confined to a very small group of workers.

In 1950 the author<sup>2</sup> reperformed Rosenfeld's self-energy calculation in a manifestly Lorentz-covariant and gauge-invariant manner. This work was stimulated by the then new "renormalization" methods, which

\* This research was supported in part by the Air Force Office of Scientific Research under Grant No. AFOSR-153-64.

† Permanent address.

<sup>1</sup> L. Rosenfeld, Ann. Physik **5**, 113 (1930); Z. Physik **65**, 589 (1930).

<sup>2</sup> B. S. DeWitt, Ph.D. thesis, Harvard University, 1950 (unpublished).

had been developed by Tomonaga, Schwinger, and Feynman, and had as its aim a demonstration that Rosenfeld's result implies merely a renormalization of charge rather than a nonvanishing photon mass. An unanticipated source of potential difficulty arose in this calculation from the fact that not one but *two* gauge groups are simultaneously present (the group associated with gravity in addition to the familiar electromagnetic group) and that these groups are not combined in the form of a direct product but rather in the form of a *semidirect product* based on the automorphisms of the electromagnetic gauge group under general coordinate transformations. This means that if a fixed choice of gauge is to be maintained, every coordinate transformation must be accompanied by an electromagnetic gauge transformation. The calculation was pushed, however, again only to the lowest order of perturbation theory; in this order, which involves only single closed Feynman loops, the ensuing complications are easily dealt with.

At about the same time investigations of a more ambitious kind were undertaken by Bergmann.<sup>3</sup> Although the renormalization philosophy had proved a resounding success in quantum electrodynamics it was still under critical attack because the methods then (and frequently even now) in use involved the explicit manipulation of divergent quantities. Similar (although more elementary) difficulties also persisted in classical particle theories with one important exception, namely, the theory of the interaction of point masses with gravity. In 1938, Einstein, Infeld, and Hoffmann<sup>4</sup> had shown that the laws of motion of such particles follow from the gravitational field equations alone, without divergent quantities ever appearing or such concepts as self-mass intervening at any time. Moreover, this result had been subsequently extended to include electrically charged particles, and gave promise of being applicable to spinning particles as well. The gravitational field thus appeared as a kind of classical regulator, and Bergmann reasoned that the same might be true in the quantum theory. Since the fields are basic, in the Einstein-Infeld-Hoffmann view, and the particles are merely singularities in the fields, Bergmann's first task was to quantize the gravitational field. It was to be hoped that commutation relations for particle position and momentum would then follow as corollaries.

The obstacles which Bergmann faced were enormous. First of all, since the laws of particle motion depend crucially on the nonlinear properties of the Einstein field equations, it was necessary to quantize the full nonlinear gravitational field. Secondly, it was necessary to find some way of defining particle position and momentum in terms of field variables alone. Thirdly, it would eventually be necessary to include spin, so that

<sup>3</sup> A bibliography of Bergmann's early work will be found in P. G. Bergmann, *Helv. Phys. Acta Suppl.* 4, 79 (1956).

<sup>4</sup> A. Einstein, L. Infeld, and B. Hoffmann, *Ann. Math.* 39, 65 (1938).

quantized particles obeying a Dirac-like equation could be described. Fourthly, it would be necessary to extract Fermi statistics (for the particles) out of the Bose statistics obeyed by the gravitational field. Finally, it would be necessary ultimately to remove the asymmetry between particle and field inherent in the Einstein-Infeld-Hoffmann approach, so as to be able, as in quantum electrodynamics, to account for pair production and vacuum polarization. It is not surprising that Bergmann's goal today remains as elusive as ever.

To achieve this goal Bergmann set out upon the classical canonical road in search of a Hamiltonian. Despite the fact that canonical methods, by singling out the time for special treatment, run counter to the spirit of any relativistic theory—above all, such a completely covariant theory as general relativity—such a procedure seemed a good one for several reasons. Firstly, no other method was then known. Secondly, canonical methods afford quick insights into certain aspects of any theory. Thirdly, it seemed that standard perturbation methods would become available for certain types of calculations.

However, Bergmann immediately ran into major difficulties (some of which had already been foreseen by Rosenfeld) in the first stages of his program. These are referred to as "the problem of constraints," and are manifested in the following ways: Some of the field variables possess no conjugate momenta; the momenta conjugate to the remaining field variables are not all dynamically independent; the field equations themselves are not linearly independent, and some of them involve no second time derivatives, thus complicating the Cauchy problem. These difficulties are all related and arise from the existence of the general coordinate-transformation group as an invariance group for the theory.

Similar difficulties had already been encountered with the electromagnetic field and methods for handling them were well known. The same methods, however, proved to be much more difficult to apply in the case the gravitational field. An obstacle is created, for example, by the fact that not all of the relations between the momenta (i.e., the constraints) are linear. Moreover, because the invariance group of gravity is non-Abelian (in contrast to the gauge group of electrodynamics) tedious calculations must be performed to check that the commutators of the various constraints lead to no inconsistencies.

Bergmann and his co-workers performed much valuable ground work in formulating the difficulties in a precise way and in partially resolving them. In the meantime additional help came from an unexpected quarter. In 1950 Dirac<sup>5</sup> published the outline of a general Hamiltonian theory which is in principle applicable to any system describable by an action functional. Dirac's methods were quickly seized upon by Pirani and Schild<sup>6</sup>

<sup>5</sup> P. A. M. Dirac, *Can. J. Math.* 2, 129 (1950).

<sup>6</sup> F. A. E. Pirani and A. Schild, *Phys. Rev.* 79, 986 (1950).

for application to the gravitational field. Unfortunately, these authors chose to develop the theory within the framework of a "parameter formalism," in the hope, which eventually proved to be misplaced, of retaining a manifest covariance which Dirac's methods would otherwise destroy. The complexity of the resulting algebra prevented them from computing all of the constraints.

The theory remained in this incomplete state for several years. It was not until impetus was provided by the first international relativity conference in Bern in 1955 (Jubilee of Relativity Theory) and the second one in Chapel Hill in 1957 that things began to move again. A small step forward was made by the author,<sup>7</sup> who showed, using the Pirani-Schild formalism, that the four so-called "primary" constraints could, by a phase transformation, be changed into pure momenta. This meant that the state functional for gravity must be independent of the  $g_{\mu\nu}$  components of the metric tensor ( $\mu=0, 1, 2, 3$ ). Shortly afterward Higgs<sup>8</sup> showed that three of the so-called "secondary" or "dynamical" constraints are the generators of infinitesimal transformations of the three "spatial" coordinates  $x^1, x^2, x^3$ . The implication of this was that the state functional must be independent of the coordinates chosen in the spacelike cross sections  $x^0=\text{constant}$  and hence cannot be taken to be an *arbitrary* functional of the metric components  $g_{ij}$  ( $i, j=1, 2, 3$ ). Developments thereafter came rapidly. Dirac himself had by this time begun to apply his methods to the gravitational field.<sup>9</sup> As a result of simplifications and clarifications which he introduced, it became easy to show that the fourth dynamical constraint is consistent with the others, and the formal theory achieved for the first time a state of technical completion. It was then possible to begin asking "What does it all mean?"

On the classical side, the problems of physical interpretation were soon resolved by the work of Arnowitt, Deser, and Misner,<sup>10</sup> who showed how to use the canonical theory to provide a rigorous characterization of gravitational radiation and "energy." In the quantum domain, however, the interpretation of the formalism remained puzzling and obscure for several years, because one did not know the right questions to ask. It is only recently that the relevant issues have begun to come into focus, largely as a result of the patient researches of Wheeler,<sup>11</sup> whose ideas have proved a great source of stimulation to many workers, including the author.

<sup>7</sup> Reported at a meeting at Stevens Institute of Technology in January, 1958 (unpublished).

<sup>8</sup> P. W. Higgs, Phys. Rev. Letters 1, 373 (1958); 3, 66 (1959).

<sup>9</sup> P. A. M. Dirac, Proc. Roy. Soc. (London) A246, 326 (1958); A246, 333 (1958); Phys. Rev. 114, 924 (1959).

<sup>10</sup> R. Arnowitt, S. Deser, and C. W. Misner, in *Gravitation: An Introduction to Current Research*, edited by L. Witten (John Wiley & Sons, Inc., New York, 1962).

<sup>11</sup> The work of Wheeler and his associates is well described in J. A. Wheeler, *Relativity Groups and Topology, 1963 Les Houches Lectures* (Gordon and Breach Science Publishers, Inc., New York, 1964). This reference contains a large bibliography of additional papers on quantization, collapse, and many other related topics.

The present paper is the direct outcome of conversations with Wheeler,<sup>12</sup> during which one fundamental question in particular kept recurring: *What is the structure of the domain manifold for the quantum-mechanical state functional?* The attempt to answer this question has required a more far-reaching analysis of the technical structure of the canonical theory than can be found in the previous literature. The results of this analysis are here presented and used to develop an interpretative framework which, although tentative, is perhaps capable of serving in a variety of contexts.

Attention is mainly confined to the case of closed finite worlds, firstly because the issues which finite worlds raise are more critical and bizarre, and secondly because the case of infinite worlds is better handled within the framework of the so-called manifestly covariant theory which will be treated in two subsequent papers of this series. The latter theory, which has also achieved a state of technical completion following the pioneering work of Feynman,<sup>13</sup> differs utterly in its structure from the canonical theory, and so far no one has established a rigorous mathematical link between the two. At the present time the two theories play complementary roles, the canonical theory describing the quantum behavior of 3-space regarded as a time-varying geometrical object, and the covariant theory describing the behavior of real and virtual gravitons propagating in this object.

Section 2 of the present paper begins with the derivation of the canonical Lagrangian. Its structure in terms of the extrinsic and intrinsic curvatures of the hypersurface  $x^0=\text{constant}$  is displayed, and attention is called to its relation to the total field energy in an asymptotically flat world. Section 3 is devoted to the primary and secondary constraints of the theory and to the independent question of coordinate conditions. Quantization is introduced in Sec. 4. Here the puzzling question of the role of a vanishing Hamiltonian is resolved by emphasizing the distinction between finite and infinite worlds. Asymptotic energy is an indispensable concept in an infinite world, and the Hamiltonian must be chosen accordingly. In a finite world there is no asymptotic energy, and an intrinsic description of the dynamics must be found, based on the constraints alone. The consistency of the constraints is demonstrated by straightforward computation of their commutators. The factor-ordering problem is disposed of by formal arguments which in effect assert that field variables taken at the same space-time point should be regarded as freely commutable. The "x constraints" are shown to be the generators of 3-dimensional coordinate transformations.

In Sec. 5 the metric representation is introduced. The

<sup>12</sup> Any errors or wrong conjectures it contains are the author's own.

<sup>13</sup> R. P. Feynman, Mimeographed letter to V. F. Weisskopf dated January 4 to February 11, 1961 (unpublished); Acta Phys. Polon. 24, 697 (1963).

distinction between finite and infinite worlds is again noted, and it is emphasized that the state functional in the former case depends only on the 3-geometry of the hypersurface  $x^0 = \text{constant}$  and not on the label  $x^0$  itself. The concept of a manifold  $\mathcal{M}$  of 3-geometries is introduced, and the role played by the Hamiltonian constraint in determining its geometrical structure is suggested. The coefficient of the momenta in the Hamiltonian constraint may be regarded as a metric of a 6-dimensional hyperbolic Riemannian manifold  $M$ . The structure of this manifold is studied in detail, and its geodesic incompleteness, owing to the existence of a frontier of infinite curvature, is noted. The possibility of relating  $M$  to the question of "intrinsic time" for the state functional is discussed, and a natural definition for the inner product of two state functionals is proposed.

In Sec. 6 a natural metric based on  $M$  is assigned to the infinite-dimensional manifold  $\mathcal{M}$ , and some of the properties of geodesics in  $\mathcal{M}$  are examined. An attempt is then made to indicate the extent to which the dynamical properties of the quantized gravitational field are determined by the structure of  $\mathcal{M}$ . The attempt is heuristic and far from complete, and much work remains to be done. The problem is approached through the WKB approximation and Hamilton-Jacobi theory. Einstein's equations are revealed as geodesic equations in  $\mathcal{M}$ , modified by the presence of a "force term." The classical phenomenon of gravitational collapse shows that the force term is not powerful enough to prevent the trajectory of 3-space from striking the frontier of  $\mathcal{M}$ . The problem of determining when the collapse phenomenon represents a real barrier to the quantum state functional is briefly discussed, and a boundary condition (vanishing state functional) at the barrier is proposed.

The barrier boundary condition raises difficulties with the definition of probability. In order to study these difficulties it is useful to test the theory on a simplified model. In Sec. 7 the quantized Friedmann universe is studied in detail, and its static wave functions in the WKB approximation are obtained. In order to obtain nonstatic wave functions which resemble a dynamical universe evolving it is necessary to introduce a clock. The combined wave functions of universe-cum-clock are studied, and it is pointed out that normalizability of the wave functions requires precise commensurability between the periods of universe and clock.

Wave packets exhibiting quasiclassical behavior are constructed in Sec. 8, in three different representations. Two of these make use of proper times defined by the clock and the universe respectively; the third treats universe and clock symmetrically through their mutual correlations. Attention is called to the deficiencies of the first two representations arising from the fact that, in a covariant theory, time is only a phenomenological concept. In the third representation probability flows in a closed finite circuit in configuration space, and wave packets do *not* ultimately spread in time. Use is made of

this fact in Sec. 9 to show how the inner-product definition can be rescued from the negative probability difficulties arising from the barrier boundary condition  $\Psi=0$  at  $R=0$  ( $R$ =radius of universe). It is also shown that the conventional Cauchy data for the wave function suffice to determine the quantum state completely.

Section 10 is devoted to speculations on the general theory. An interpretation of quantum mechanics due to Everett (see Ref. 52) is described and proposed for dealing with the concept of "a wave function for the universe." Such an interpretation is essential if the wave function is unique. Evidence is presented that the Hamiltonian constraint may indeed have only one solution. The problem of time-reversal invariance and entropy is briefly discussed. Two technical appendices follow at the end of the article.

Attention is called to the following points of notation: Latin indices range over the values 1, 2, 3 and Greek indices over the values 0, 1, 2, 3. Differentiation is denoted by a comma. The coordinates  $x^0$  and  $x^i$  are assumed to be timelike and spacelike, respectively, and the geometry of space-time is assumed to be such that the hypersurfaces  $x^0 = \text{constant}$  are capable of carrying a complete set of Cauchy data. So-called "absolute units" in which  $\hbar=c=16\pi G=1$  ( $G$  being the gravitation constant) are used throughout, as is also the signature  $-+++$  for the space-time metric  $g_{\mu\nu}$ . The Riemann and Ricci tensors, and the curvature scalar, are taken in the respective forms

$$R_{\mu\nu\sigma}{}^\tau = \Gamma_{\nu\sigma}{}^\tau,_\mu - \Gamma_{\mu\sigma}{}^\tau,_\nu + \Gamma_{\nu\sigma}{}^\rho \Gamma_{\mu\rho}{}^\tau - \Gamma_{\mu\sigma}{}^\rho \Gamma_{\nu\rho}{}^\tau, \quad (1.1)$$

$$R_{\mu\nu} \equiv R_{\sigma\mu\nu}{}^\sigma, \quad (1.2)$$

$${}^{(4)}R \equiv R_{\mu}{}^{\mu} \equiv g^{\mu\nu} R_{\mu\nu}, \quad (1.3)$$

where

$$\Gamma_{\mu\nu}{}^\sigma \equiv \frac{1}{2} g^{\sigma\tau} (g_{\mu\tau,\nu} + g_{\nu\tau,\mu} - g_{\mu\nu,\tau}), \quad g_{\mu\sigma} g^{\sigma\nu} = \delta_\mu{}^\nu. \quad (1.4)$$

The corresponding tensors in the spacelike cross sections  $x^0 = \text{constant}$  are distinguished by means of a prefixed superscript (3). These conventions have the property that  ${}^{(4)}R$  is non-negative in a space-time containing normal matter and satisfying Einstein's equations, and that  ${}^{(3)}R$  is positive in a 3-space of positive curvature.

## 2. EXTRINSIC AND INTRINSIC CURVATURE. CLASSIC FORM OF THE LAGRANGIAN

The canonical theory begins with the following decomposition of the metric tensor:

$$(g_{\mu\nu}) = \begin{pmatrix} -\alpha^2 + \beta_i \beta^i & \beta_j \\ \beta_i & \gamma_{ij} \end{pmatrix}, \quad (2.1)$$

$$(g^{\mu\nu}) = \begin{pmatrix} -\alpha^{-2} & \alpha^{-2} \beta^j \\ \alpha^{-2} \beta^i & \gamma^{ij} - \alpha^{-2} \beta^i \beta^j \end{pmatrix},$$

$$\gamma_{ik} \gamma^{kj} = \delta_i{}^j, \quad \beta^i = \gamma^{ij} \beta_j. \quad (2.2)$$

When the conventional Einstein Lagrangian density is reexpressed in terms of the new variables, it is found, after some calculation, to take the form

$$\mathcal{L} \equiv g^{1/2} {}^{(4)}R = \alpha\gamma^{1/2}(K_{ij}K^{ij} - K^2 + {}^{(3)}R) - 2(\gamma^{1/2}K)_{,0} + 2(\gamma^{1/2}K\beta^i - \gamma^{1/2}\gamma^{ij}\alpha_{,j}), \quad (2.3)$$

where

$$g \equiv -\det(g_{\mu\nu}) = \alpha^2\gamma, \quad \gamma \equiv \det(\gamma_{ij}), \quad (2.4)$$

$$K_{ij} \equiv \frac{1}{2}\alpha^{-1}(\beta_{i,j} + \beta_{j,i} - \gamma_{ij,0}), \quad K^{ij} \equiv \gamma^{ik}\gamma^{jl}K_{kl}, \quad (2.5)$$

$$K \equiv \gamma^{ij}K_{ij},$$

the dots denoting covariant differentiation based on the 3-metric  $\gamma_{ij}$ .

The quantity  $K_{ij}$ , which transforms as a symmetric tensor under spatial coordinate transformations, is known as the *second fundamental form*. It describes the curvature of the hypersurface  $x^0 = \text{constant}$  as viewed from the 4-dimensional space-time in which it is embedded. It is therefore also frequently called the *extrinsic curvature tensor* of the hypersurface, as opposed to the *intrinsic curvature tensor*  ${}^{(3)}R_{ij}$ , which depends only on  $\gamma_{ij}$  in the hypersurface. In a flat space-time  ${}^{(3)}R_{ij}$  is completely determined by  $K_{ij}$ , but in a manifold of arbitrary curvature there need be no relationship between the two. The contracted forms  ${}^{(3)}R$  and  $K_{ij}K^{ij} - K^2$  will be referred to as the intrinsic and extrinsic curvatures, respectively.

The last two terms of Eq. (2.3), being total derivatives, are dynamically irrelevant and may be dropped. The Lagrangian then becomes

$$L \equiv \int \alpha\gamma^{1/2}(K_{ij}K^{ij} - K^2 + {}^{(3)}R)d^3x, \quad (2.6)$$

which has the classic form “kinetic energy minus potential energy,” with the extrinsic curvature playing the role of kinetic energy and the negative of the intrinsic curvature that of potential energy.

The form (2.6) is manifestly invariant under 3-dimensional general coordinate transformations. Precisely for this reason it differs from the Lagrangian of ordinary field theories, for the  ${}^{(3)}R$  term of its integrand contains linearly occurring second spatial derivatives of the field variables. With an ordinary field theory in an infinite universe this would be of no significance. The usual assumption that the field vanishes outside some arbitrarily large but finite spatial domain permits linearly occurring second derivatives to be eliminated by partial integration without affecting either the dynamical equations or the canonical definition of energy. In the case of gravity, however, the field never vanishes outside a finite domain unless space-time is flat, and although such a partial integration leaves the dynamical equations unaffected it does change the definition of energy. It is easy to verify, in fact, that it subtracts

from the Lagrangian (2.6) a surface integral  $E_\infty$  given by

$$E_\infty = \int_{\infty} \alpha\gamma^{1/2}\gamma^{ij}(\gamma_{ik,j} - \gamma_{ij,k})dS^k, \quad (2.7)$$

and hence adds a corresponding quantity to the canonical energy. In an asymptotically flat world it is always possible to find an asymptotically Minkowskian reference frame in which  $\alpha$ ,  $\beta_i$ , and  $\gamma_{ij}$  take the static Schwarzschild forms

$$\alpha \rightarrow 1 - \frac{M}{16\pi r}, \quad \beta_i \rightarrow 0, \quad \gamma_{ij} \rightarrow \delta_{ij} + \frac{M}{8\pi} \frac{x^ix^j}{r^3}, \quad (2.8)$$

where  $r^2 \equiv x^ix^i$  and  $M$  is the effective gravitational mass of the field distribution. Substitution of (2.8) into (2.7) yields

$$E_\infty = M. \quad (2.9)$$

It is to be noted that the removal of  $E_\infty$  from the Lagrangian does not correspond to a mere redefinition of the energy zero point.  $E_\infty$  is not a fixed constant but depends on the state of the field. In fact it is the energy, for as we shall see presently the canonical “energy” based on (2.6) always vanishes. (Indeed,  $E_\infty$  is the energy even when other fields are present.) Since neither (2.6) nor (2.7) have any explicit dependence on  $x^0$ , the quantity  $E_\infty$  is conserved. General relativity is unique among field theories in that its energy may always be expressed as a surface integral. This was the source of Bergmann’s hope to use gravity as a regulator, but it is also a source of difficulties. We note in particular that the surface integral vanishes for a closed finite world. It is only for infinite asymptotically flat worlds that the energy concept has meaning.

### 3. THE CONSTRAINTS

The momenta conjugate to  $\alpha, \beta_i$ , and  $\gamma_{ij}$  will be denoted by  $\pi$ ,  $\pi^i$ , and  $\pi^{ij}$ , respectively. They have the explicit forms

$$\pi = \frac{\delta L}{\delta \dot{\alpha}_{,0}} = 0, \quad (3.1)$$

$$\pi^i = \frac{\delta L}{\delta \dot{\beta}_{i,0}} = 0, \quad (3.2)$$

$$\pi^{ij} = \frac{\delta L}{\delta \dot{\gamma}_{ij,0}} = -\gamma^{1/2}(K^{ij} - \gamma^{ij}K), \quad (3.3)$$

Eqs. (3.1) and (3.2) being known as the *primary constraints*. The primary constraints are purely formal statements, which express the fact that the Lagrangian (2.6) is independent of the “velocities”  $\alpha_{,0}$  and  $\beta_{i,0}$ .<sup>14</sup>

<sup>14</sup> Failure to bring the Lagrangian into the form (2.6) was responsible for the difficulties originally encountered with the primary constraints.

These "velocities" are arbitrary and cannot be re-expressed in terms of momenta. They therefore cannot be removed from the Hamiltonian, which, with the aid of (3.3), takes the form

$$\begin{aligned} H &= \int (\pi\alpha_{,0} + \pi^i\beta_{i,0} + \pi^{ij}\gamma_{ij,0}) d^3x - L \\ &= \int (\pi\alpha_{,0} + \pi^i\beta_{i,0} + \alpha\mathfrak{C} + \beta_i\chi^i) d^3x, \end{aligned} \quad (3.4)$$

where

$$\mathfrak{C} \equiv \frac{1}{2}\gamma^{-1/2}(\gamma_{ik}\gamma_{jl} + \gamma_{il}\gamma_{jk} - \gamma_{ij}\gamma_{kl})\pi^{ij}\pi^{kl} - \gamma^{1/2}(3)R \quad (3.5a)$$

$$= \gamma^{1/2}(K_{ij}K^{ij} - K^2 - (3)R), \quad (3.5b)$$

$$\chi^i \equiv -2\pi^{ij,j} \equiv -2\pi^{ij,j} - \gamma^{il}(2\gamma_{jl,k} - \gamma_{jk,l})\pi^{jk}. \quad (3.6)$$

The momenta, as well as  $\mathfrak{C}$  and  $\chi^i$ , are all 3-densities of unit weight.

It is not hard to show that Einstein's empty-space field equations may be obtained by taking the Poisson bracket of the various dynamical variables with the Hamiltonian (3.4) and then imposing Eqs. (3.1) and (3.2) as external constraints. Since the undermined "velocities"  $\alpha_{,0}$  and  $\beta_{i,0}$  are multiplied in (3.4) by  $\pi$  and  $\pi^i$ , their Poisson brackets with anything may be ignored. If desired, one can always assign definite values to  $\alpha$  and  $\beta_i$  which may be purely numerical or may depend on the  $\gamma_{ij}$  and  $\pi^{ij}$ . Each choice corresponds to the imposition of certain conditions on the space-time coordinates. For example, one may choose

$$\alpha = 1, \quad \beta_i = 0, \quad (3.7a)$$

which reduces the Hamiltonian to

$$H = \int \mathfrak{C} d^3x. \quad (3.7b)$$

Another favorite choice is

$$K \equiv \frac{1}{2}\gamma^{-1/2}\gamma_{ij}\pi^{ij} = 0, \quad (\gamma^{1/2}\gamma^{ij})_{,j} = 0, \quad (3.8a)$$

which corresponds to the requirement that the volume of every hypersurface  $x^0 = \text{constant}$  be stationary under small timelike deformations,<sup>15</sup> and that the spatial coordinates in each hypersurface be harmonic. To obtain the explicit forms of the conditions which Eqs. (3.8a) impose upon  $\alpha$  and  $\beta_i$ , one notes that these equations imply the vanishing not only of their left-hand sides but of all their space-time derivatives as well. Taking the Poisson brackets of  $K$  and  $(\gamma^{1/2}\gamma^{ij})_{,j}$  with the Hamiltonian (3.4) one finds the conditions

$$\begin{aligned} \alpha_{,i} - (3)R\alpha &= 0, \\ [\gamma^{1/2}(\beta_{i,j} + \beta_{j,i} - \gamma^{ij}\beta_{k,k}) + 2\alpha\pi^{ij}]_{,j} &= 0. \end{aligned} \quad (3.8b)$$

In an infinite asymptotically flat world these equations, which are of the elliptic type, may be solved subject to

<sup>15</sup> If 3-space is infinite this applies to the volume inside every finite domain.

the boundary conditions  $\alpha \rightarrow 1$ ,  $\beta_i \rightarrow 0$  at infinity in asymptotically Minkowskian coordinates. In a finite world of nonvanishing curvature, however, they usually possess either no physically admissible solutions, i.e., solutions for which  $\alpha$  remains everywhere positive, or no solutions at all. Since the Laplace-Beltrami operator has a negative spectrum, the first equation, for example, cannot be solved in a 3-sphere.

Conditions of the above type correspond merely to restrictions on the coordinates and have no physical content. There exist conditions of yet another type which actually restrict the dynamical freedom of the field and which hold regardless of whether a specific choice has been made for  $\alpha$  and  $\beta_i$  or not. These are obtained by noting that since the primary constraints hold for all time, the  $x^0$  derivatives of  $\pi$  and  $\pi^i$  must vanish. Stating this in the form of a Poisson bracket with  $H$ , one arrives immediately at the so-called *secondary or dynamical constraints*:

$$\mathfrak{C} = 0, \quad (3.9)$$

$$\chi^i = 0. \quad (3.10)$$

Equation (3.9) will be called the *Hamiltonian constraint*<sup>16</sup> in virtue of the structure of the function  $\mathfrak{C}$ , which appears in (3.5b) as the difference of the extrinsic and intrinsic curvatures, in analogy with the classic form of the Hamiltonian as the sum of the kinetic and potential energies. This Hamiltonian, however, vanishes, as does indeed the total Hamiltonian (3.4). That is to say, in any "Ricci-flat" space-time (i.e., one satisfying Einstein's empty-space equations) the extrinsic and intrinsic curvatures of any hypersurface are equal. As has been emphasized by Wheeler,<sup>11</sup> the converse of this theorem is also true, namely, if  $\mathfrak{C}$  vanishes over every hypersurface then space-time is Ricci-flat. This suggests, as will be verified later, that it is the Hamiltonian constraint which provides the essential description of the "intrinsic" (i.e., coordinate-independent) dynamics of the gravitational field.

#### 4. QUANTIZATION, CONSISTENCY OF THE CONSTRAINTS, FACTOR ORDERING

In the quantum theory, Poisson brackets become commutators. This means that the constraint Eqs. (3.1), (3.2), (3.9), and (3.10) cannot become operator equations, for otherwise the Hamiltonian (3.4) would yield no dynamics at all, extrinsic or intrinsic. Instead they become conditions on the state vector  $\Psi^{5,9}$ :

$$\pi\Psi = 0, \quad (4.1)$$

$$\pi^i\Psi = 0, \quad (4.2)$$

$$\mathfrak{C}\Psi = 0, \quad (4.3)$$

$$\chi^i\Psi = 0. \quad (4.4)$$

<sup>16</sup> All four constraints (3.9), (3.10) are sometimes referred to as "Hamiltonian constraints." We prefer to reserve the terminology for this particularly important constraint.

These quantum constraints are often a source of puzzlement and confusion. Consider the equation

$$\gamma_{ij}(x^0, \mathbf{x}) = e^{iHx^0} \gamma_{ij}(0, \mathbf{x}) e^{-iH^0}, \quad \mathbf{x} = (x^1, x^2, x^3) \quad (4.5)$$

which is the quantum-mechanical relation expressing the field operator  $\gamma_{ij}$  on an arbitrary hypersurface in terms of the corresponding operator on the hypersurface  $x^0=0$ . Suppose we choose  $\alpha$  and  $\beta_i$  as in Eq. (3.7a). Then Eq. (4.3) and its conjugate imply<sup>17</sup>

$$H\Psi = 0, \quad \Psi^\dagger H = 0, \quad (4.6)$$

and hence

$$\Psi^\dagger \gamma_{ij}(x^0, \mathbf{x}) \Psi = \Psi^\dagger \gamma_{ij}(0, \mathbf{x}) \Psi. \quad (4.7)$$

A similar result holds for any other field operator or product of field operators. Since the statistical results of any set of observations are ultimately expressible in terms of expectation values, one therefore comes to the conclusion that nothing ever happens in quantum gravodynamics, that the quantum theory can never yield anything but a static picture of the world.<sup>18</sup>

To see what is wrong with this conclusion one must examine the behavior of  $H$ , or more precisely  $\mathcal{H}$ , at infinity. In an infinite asymptotically flat world the field disperses ultimately to a state of infinite weakness. In the asymptotic region  $\mathcal{H}$  therefore tends to its dominant linear term  $\gamma_{ii,ij} - \gamma_{ij,ij} = 0$ , which is the well-known fourth constraint of linearized gravity theory.<sup>19</sup> This term is the asymptotic limit of the term which is removed from the integrand of  $H$  by the partial integration discussed in Sec. 2, and which gives rise to the surface integral (2.7). In the linearized theory, however, it becomes a constraint which has no relation to the total energy. Therefore if the full theory is to be applicable not only in the nonlinear region but also at infinity where the linear theory holds sway, it must make use of the Hamiltonian

$$H_\infty \equiv H + E_\infty, \quad (4.8)$$

which results from the partial integration. The integrand of this Hamiltonian reduces, in the asymptotic region, to an expression quadratic in the  $\gamma$ 's and  $\pi$ 's, namely, the usual integrand of the linearized theory.

It follows that in an infinite asymptotically flat world Eq. (4.5) should be replaced by

$$\gamma_{ij}(x^0, \mathbf{x}) = e^{iH_\infty x^0} \gamma_{ij}(0, \mathbf{x}) e^{-iH_\infty x^0}. \quad (4.9)$$

Even with this replacement, however, the appearance of the world is still static whenever  $\Psi$  is an eigenstate of energy-momentum. To obtain nonstatic behavior one must construct *wave packets*, by superposing many different momenta. But this is precisely what one wants to do in order to provide *S*-matrix theory, for example, with a rigorous foundation and insure that the field really does disperse ultimately to a state of infinite weakness.

<sup>17</sup>  $\mathcal{H}$  is assumed to be ordered in an Hermitian fashion.

<sup>18</sup> Cf. A. Komar, Phys. Rev. **153**, 1385 (1967).

Although the above discussion makes use of the coordinate system defined by Eqs. (3.7a), the same problems arise in any other asymptotically Minkowskian coordinate system, and the same conclusions apply. To the extent that we can ignore the possible lack of commutativity of  $\alpha$  and  $\beta_i$  with  $\mathcal{H}$  and  $\chi^i$  in the construction of an Hermitian Hamiltonian, the same apparent static behavior of the field will occur whenever we incorrectly use  $H$  instead of  $H_\infty$  in Eq. (4.9).

It should be noted that coordinate conditions such as (3.7a) and (3.8a) are *operator* equations and not constraints on the state vector.<sup>19</sup> (This follows from the complete arbitrariness of  $\alpha$  and  $\beta_i$  in the classical theory.) On the other hand, equations such as (3.8a), which hold only when  $\alpha$  and  $\beta_i$  are suitably restricted, are *not* operator equations. Indeed, they are not even constraints, but become instead *expectation-value equations*

$$\Psi^\dagger K \Psi = 0, \quad \Psi^\dagger (\gamma^{1/2} \gamma^{ij})_{,j} \Psi = 0, \quad (3.8c)$$

which hold for all values of  $x^0$  provided they hold at some initial instant and Eqs. (3.8b) are satisfied. They do not hold in all permissible states merely in virtue of (3.8b).

Although we know that the physical content of the classical theory is unaffected by the choice of coordinates, it is not so easy to prove, using the canonical theory, that the results of a calculation of some physical *quantum* amplitude is independent of the choice of coordinates. It is not enough merely to know, for example, that two different coordinate systems both take the Minkowskian form  $\alpha \rightarrow 1$ ,  $\beta_i \rightarrow 0$ ,  $\gamma_{ij} \rightarrow \delta_{ij}$  at infinity, in order to conclude that the physical *S* matrix remains unchanged under the transformation from one system to the other, for the operator  $\Delta H$ , which represents the change in the Hamiltonian in passing from one system to the other, produces effects which propagate to infinity. In order to prove invariance of the *S* matrix under coordinate transformations (including *q*-number coordinate transformations), one would have to show that  $\Delta H$  affects only the nonphysical field modes at infinity. The obstacle to such a demonstration is the lack of commutativity of the operators appearing in the dynamical equations, particularly when  $\alpha$  and  $\beta_i$  depend nonlocally on  $\gamma_{ij}$  and  $\pi^{ij}$ . Although noncommutativity has no effect on the scattering amplitudes in lowest order, it plays havoc with the radiative corrections. For the study of radiative corrections a manifestly covariant theory is almost essential. In the following paper of this series the theory of gravitational radiative

<sup>19</sup> There is an alternative approach to the quantum theory of gravity which makes use of an action functional which is not coordinate-invariant and which generates no primary or secondary constraints. In this approach the constraints must be imposed from the outside. They take the form of coordinate conditions whose form is not arbitrary but is determined by the action functional itself. In this case the coordinate conditions are constraints on the state vector. This is the approach which has been followed, for example, by Gupta [S. N. Gupta, in *Recent Developments in General Relativity* (Pergamon Press, Inc., New York, 1962)].

corrections will be displayed in all its complexity, and the  $S$  matrix in the manifestly covariant theory will be proved to be fully coordinate invariant. This result has not yet been proved in the canonical theory, and for this reason we shall include little further discussion of the case of infinite asymptotically flat worlds in this paper, but will concentrate henceforth on finite worlds.

In the finite case there is no distinction between  $H$  and  $H_\infty$ , and hence we must face up anew to the difficulties posed by Eq. (4.7). The following procedure will be adopted: Instead of regarding this equation as implying that the universe is static we shall interpret it as informing us that the coordinate labels  $x^\mu$  are really irrelevant. Physical significance can be ascribed only to the intrinsic dynamics of the world, and for the description of this we need some kind of intrinsic coordinatization based either on the geometry or the contents of the universe. In the case of infinite asymptotically flat worlds the Minkowski coordinates at infinity have independent physical relevance as preferred coordinates (up to a Lorentz transformation) based on an *a priori assumed* isometry group (the Poincaré group) for the asymptotic region. One may say that they are intrinsically determined by an implicit laboratory or observer at infinity, and that the constraints serve merely to eliminate the nonphysical modes from the field. In the case of finite worlds, however, the constraints are everything; they and they alone must yield the complete quantum-mechanical description of the world geometry. One of our tasks in the remainder of this paper will be to try to convince the reader that the equations of constraint really do *saturate* the theory, that nothing else is needed.

We must first establish the fact that the constraints are consistent with each other, and this raises some issues of factor ordering.<sup>20</sup> Unfortunately, general agreement has not yet been reached on how to resolve these issues, and hence the proposals which follow must be regarded as tentative. We emphasize, however, our view that the factor-ordering question is not very important to the theory as a whole, and should in no case be permitted to impede attempts to apply the theory to concrete problems. It arises in every local-field theory possessing nontrivial spectral functions, and bears mainly on problems of interpreting divergences. The latter are always resolved by symmetry arguments or by removing infinities from divergent integrals in an invariant way. How such procedures operate in the case of gravity will appear in the papers devoted to the manifestly covariant theory, where questions of factor ordering will again be discussed.

Consistency of the constraints is established if it can be shown that commutators of the constraints lead to no new constraints. The basic commutation relations

<sup>20</sup> See, for example, J. L. Anderson, in *Proceedings of the 1962 Eastern Theoretical Conference*, edited by M. E. Rose (Gordon and Breach Science Publishers, Inc., New York, 1963), p. 387. See also J. Schwinger, Phys. Rev. 130, 1253 (1963); 132, 1317 (1963).

of the canonical variables themselves are

$$[\alpha, \pi'] = i\delta(\mathbf{x}, \mathbf{x}'), \quad [\beta_i, \pi^{i'}] = i\delta_i{}^{i'}, \quad [\gamma_{ij}, \pi^{k' l'}] = i\delta_{ij}{}^{k' l'}, \quad (4.10)$$

all other commutators vanish;

in which a notation is employed which emphasizes the bitensor transformation character of the quantities on the right, with primes being used, either on indices or on the variables themselves, to distinguish different points of 3-space. Here  $\delta(\mathbf{x}, \mathbf{x}')$  denotes the 3-dimensional  $\delta$  function, and

$$\delta_i{}^{j'} \equiv \delta_i{}^j \delta(\mathbf{x}, \mathbf{x}'), \quad \delta_{ij}{}^{k' l'} \equiv \delta_{ij}{}^{kl} \delta(\mathbf{x}, \mathbf{x}'), \quad (4.11)$$

$$\delta_{ij}{}^{kl} \equiv \frac{1}{2} (\delta_i{}^k \delta_j{}^l + \delta_i{}^l \delta_j{}^k).$$

(The  $\delta$  function will ordinarily be viewed as a bidensity of zero weight at its first argument and of unit weight at its second.)

The primary constraints evidently give no trouble, since they commute with each other and with the secondary constraints. We therefore turn to the latter and look first at the  $\chi$  constraints. These will be taken precisely as written in Eq. (3.6), with the momentum factor  $\pi^{jk}$  standing to the right. However, the index will be lowered by defining

$$\chi_i \equiv \gamma_{ij} \chi^j, \quad (4.12)$$

which, since  $\gamma_{ij}$  stands to the left, yields an alternative form for Eq. (4.4):

$$\chi_i \Psi = 0. \quad (4.13)$$

$\chi_i$  has the important property of being homogeneous bilinear in the  $\gamma_{ij}$  and the  $\pi^{ij}$ , with the  $\gamma$ 's to the left and the  $\pi$ 's to the right. Therefore its commutator with any other  $\chi_j$ , has the same property. To compute this commutator it is helpful first to compute the following:

$$\left[ \gamma_{ij}, i \int \chi_{k'} \delta \xi^{k'} d^3 x' \right]$$

$$= -\gamma_{ij,k} \delta \xi^k - \gamma_{kj,i} \delta \xi^k + \gamma_{ik,j} \delta \xi^k, \quad (4.14)$$

$$\left[ \pi^{ij}, i \int \chi_{k'} \delta \xi^{k'} d^3 x' \right]$$

$$= -(\pi^{ij} \delta \xi^k)_{,k} + \pi^{ki} \delta \xi^j_{,k} + \pi^{jk} \delta \xi^i_{,k}, \quad (4.15)$$

which reveal the  $\chi$ 's as generators of 3-dimensional coordinate transformations. Under the infinitesimal coordinate transformation  $\tilde{x}^i = x^i + \delta \xi^i$ , the change in any function of the  $\gamma_{ij}$ ,  $\pi^{ij}$  and their derivatives is given by commutation with  $i \int \chi_i \delta \xi^i d^3 x$ , provided the function has no explicit dependence on  $x$ . From this it follows at once that

$$[\chi_i, \chi_{j'}] = -i \int \chi_{k''} \delta^{k''}{}_{ij'} d^3 x'', \quad (4.16)$$

where the  $c$ 's are the structure constants of the general coordinate transformation group:

$$c^{k''}{}_{ij'} \equiv \delta^{k''}{}_{i,l'} \delta^{l'}{}_{j'} - \delta^{k''}{}_{j',l'} \delta^{l'}{}_{i'}. \quad (4.17)$$

The same observations, combined with the fact that  $\mathcal{H}$  is a scalar density, yield the formula

$$[\chi_i, \mathcal{H}'] = i\mathcal{H}\delta_{,i}(\mathbf{x}, \mathbf{x}'). \quad (4.18)$$

Again the ordering of factors remains the same on both sides of the equation. The only term of  $\mathcal{H}$  which might lead to difficulty is the one quadratic in the momenta. But all of the factors which appear in this term have homogeneous linear transformation laws under the 3-dimensional coordinate transformation group and hence remain undisturbed in position when commuted with  $\chi_i$ . Thus, the commutators (4.16) and (4.18) yield no new constraints, and the choice of factor ordering for  $\mathcal{H}$  is so far arbitrary.

Now note that all of the above results could have been obtained equally well had the opposite ordering been chosen for  $\chi_i$ , with the  $\pi$ 's standing to the left and the  $\gamma$ 's to the right. The difference between the two choices for  $\chi_i$  therefore commutes with everything and is evidently a  $c$  number. It is a  $c$  number, moreover, with definite transformation properties; namely, it is a covariant 3-vector density. From this we may conclude that it can only be zero, for otherwise 3-space would contain a preferred direction quite independently of any geometry which may be imposed on it. The reasonableness of this conclusion also follows from a straightforward formal computation of the difference between the two  $\chi$ 's, which yields derivatives of  $\delta$  functions with coincident arguments. Any ordering may therefore be chosen for  $\chi_i$ , and if  $\gamma_{ij}$  and  $\pi^{ij}$  are Hermitian so is  $\chi_i$ .

The same conclusions do not automatically hold for  $\chi^i$ , since the difference between two orderings for it involves an undifferentiated  $\delta$  function. Let us therefore see what we can say about the formal symbol  $\delta(\mathbf{x}, \mathbf{x})$ . Consider the third commutator in (4.10). If we set  $\mathbf{x}' = \mathbf{x}$  and contract all the indices, we obtain

$$(\gamma_{ij}\pi^{ij} - \pi^{ij}\gamma_{ij}) = 6i\delta(\mathbf{x}, \mathbf{x}). \quad (4.19)$$

The quantity on the right is certainly a  $c$  number. Therefore we may write

$$[6i\delta(\mathbf{x}, \mathbf{x}), i \int \chi_{k'} \delta\xi^{k'} d^3x'] = 0. \quad (4.20)$$

On the other hand, if we apply the same commutator to the left we obtain

$$\begin{aligned} & [(\gamma_{ij}\pi^{ij} - \pi^{ij}\gamma_{ij}), i \int \chi_{k'} \delta\xi^{k'} d^3x'] \\ &= -[(\gamma_{ij}\pi^{ij} - \pi^{ij}\gamma_{ij}) \delta\xi^k]_{,k} = -6i[\delta(\mathbf{x}, \mathbf{x}) \delta\xi^k]_{,k}. \end{aligned} \quad (4.21)$$

Equating the two results we find

$$[\delta(\mathbf{x}, \mathbf{x}) \delta\xi^i]_{,i} = 0. \quad (4.22)$$

This equation must hold for arbitrary  $\delta\xi^i$ . Therefore, although most people would say that  $\delta(\mathbf{x}, \mathbf{x})$  is infinite, we see that it is actually zero.

In order to understand how this formal result can be consistent with the rest of the theory one must first note that Eqs. (4.3) and (4.13) are really abbreviations for the correct forms

$$\int \mathcal{H}\xi d^3x \Psi = 0 \quad \text{for all } \xi, \quad (4.23)$$

$$\int \chi_i \xi^i d^3x \Psi = 0 \quad \text{for all } \xi^i, \quad (4.24)$$

where  $\xi$  and  $\xi^i$  are arbitrary but smooth  $c$ -number weight functions. The problem of taking commutators of field quantities at the same space-time point therefore never arises with pairs of constraints but only in connection with the definition of the functions  $\mathcal{H}$  and  $\chi_i$  themselves. This means that the  $\delta$  function may, without inconsistency, be thought of as the limit of a sequence of successively narrower *twin-peaked* functions, all of which are smooth, have unit integral, and vanish at the point  $\mathbf{x}' = \mathbf{x}$  in the valley between the peaks. An example of such a function in one dimension would be  $\delta(x) = \lim(2\pi)^{-1}[f_\epsilon(x - \sqrt{\epsilon}) + f_\epsilon(x + \sqrt{\epsilon}) - 2f_\epsilon(x)/(1+\epsilon)]$ , where  $f_\epsilon(x) \equiv \epsilon(x^2 + \epsilon^2)^{-1}$ . In an infinite world, passage to the limit  $\epsilon \rightarrow 0$  would correspond to the usual cutoff going to infinity in momentum space, while maintenance of the valley at  $\mathbf{x}' = \mathbf{x}$  would yield a particular regularization of the resulting divergences. The answer to the question whether or not this regularization is equivalent to the quite different procedures which will prove useful in the manifestly covariant theory must await a demonstration of how to derive one theory from the other. In the meantime we shall in *this* paper simply adopt it as a rule that any two field operators taken at the same space-time point commute. The consistency question for the constraints then reduces to that of the classical theory.

There remains to be considered only the commutator  $[\mathcal{H}, \mathcal{H}']$ . At first sight it might be thought that the commutator of the two quadratic-in-the-momenta terms, one from  $\mathcal{H}$  and the other from  $\mathcal{H}'$ , leads to difficulties. However, these terms contain no derivatives (of the  $\gamma$ 's or  $\pi$ 's) with respect to the 3-space coordinates and hence they commute. Since the terms  $\gamma^{1/2} {}^{(3)}R$  and  $\gamma'^{1/2} {}^{(3)}R'$  contain no momenta, they likewise commute. The only commutators which remain are the cross commutators, and these can be evaluated by judicious use of the variational formula

$$\begin{aligned} \delta(\gamma^{1/2} {}^{(3)}R) &= \gamma^{1/2} \gamma^{ij} \gamma^{kl} (\delta\gamma_{ik,jl} - \delta\gamma_{ij,kl}) \\ &\quad - \gamma^{1/2} ({}^{(3)}R^{ij} - \frac{1}{2}\gamma^{ij} {}^{(3)}R) \delta\gamma_{ij}. \end{aligned} \quad (4.25)$$

The final result is

$$[\mathcal{H}, \mathcal{H}'] = 2i\chi^i \delta_{,i}(\mathbf{x}, \mathbf{x}') + i\chi^i_{,i} \delta(\mathbf{x}, \mathbf{x}'), \quad (4.26a)$$

or, more correctly,

$$\left[ \int \Im \xi_1 d^3x, \int \Im \xi_2 d^3x \right] = i \int x^i (\xi_{1,i} - \xi_{2,i}) d^3x. \quad (4.26b)$$

If we were still concerned about the order of factors we would find that a symmetric (Hermitian) ordering for  $\Im \xi$  would yield a symmetric ordering for  $x^i$  in (4.26), namely  $x^i = \frac{1}{2}\{\gamma^{ij}, x_j\}$ , and the problem at issue would then be to evaluate the commutator  $[\gamma^{ij}, x_j]$ . From our present point of view this commutator vanishes, and consistency is maintained.<sup>21</sup>

## 5. THE FUNCTIONAL WAVE EQUATION AND THE STATE-FUNCTIONAL DOMAIN MANIFOLD

Further analysis of the canonical theory requires the introduction of a specific representation for the quantum states. Wheeler<sup>11</sup> has chosen for this purpose what may be called the *metric representation*, in which  $\Psi$  becomes a functional of the metric components  $g_{\mu\nu}$ , and the momenta become functional differential operators:

$$\pi = \frac{\delta}{i\partial\alpha}, \quad \pi^i = \frac{\delta}{i\partial\beta_i}, \quad \pi^{ij} = \frac{\delta}{i\partial\gamma_{ij}}. \quad (5.1)$$

The primary constraints tell us that Wheeler's  $\Psi$  depends only on the  $\gamma$ 's. We shall indicate this, for the present, by writing  $\Psi$  in the form  $\Psi[\gamma]$ . (Since we are working in a closed finite world, it would be meaningless to include also a dependence on  $x^0$ .)

Consider now the  $\chi$  constraints. In the metric representation these take the form

$$2i(\delta\Psi[\gamma]/\delta\gamma_{ij}).j = 0, \quad (5.2)$$

which are the necessary and sufficient conditions that  $\Psi[\gamma]$  be an invariant under coordinate transformations. In a finite world this means that  $\Psi$  depends only on the geometry of 3-space. One possible way to express this dependence would be to regard  $\Psi$  as a function of a discrete infinity of variables, namely all the independent invariants, beginning with  $\int \gamma^{1/2} d^3x$ ,  $\int \gamma^{1/2} {}^{(3)}R d^3x$ ,

<sup>21</sup> J. Schwinger (Ref. 20) proposes an alternative resolution of the factor ordering problem which, in the notation of the present paper, runs essentially as follows: Replace  $\Im \xi$  in the Hamiltonian constraint by  $(\gamma^{3/2}\Im \xi)$ , where  $( )$  indicates that the factors are to be placed in some (arbitrary) symmetrical order. Then compute

$$\begin{aligned} [(\gamma^{3/2}\Im \xi), (\gamma'^{3/2}\Im \xi')] &= \frac{1}{2}i[\{\gamma^3\gamma^{ij}, (x_j)\} + \{\gamma'^3\gamma^{ij'}, (x_{j'})\}] \delta_{ij}(x, x') \\ &= \frac{1}{2}i[\{\gamma^3\gamma^{ij}, (x_j)\} + \{\gamma'^3\gamma^{ij'}, (x_{j'})\}] \delta_{ij}(x, x'). \end{aligned}$$

Since the commutator

$$[\gamma^3\gamma^{ij}, (x_j)] = i(\gamma^3\gamma^{ij} + \gamma'^3\gamma^{ij'})\delta_{ij}(x, x')$$

is antisymmetric in  $x$  and  $x'$ , it follows that

$$[\gamma^3\gamma^{ij}, (x_j)] + [\gamma'^3\gamma^{ij'}, (x_{j'})] = 0,$$

whence

$[(\gamma^{3/2}\Im \xi), (\gamma'^{3/2}\Im \xi')] = i[\gamma^3\gamma^{ij}, (x_j) + \gamma'^3\gamma^{ij'}, (x_{j'})]\delta_{ij}(x, x')$  in which the  $\chi$ 's stand to the right. Demonstration of consistency of the other commutators is elementary.

$\int \gamma^{1/2} {}^{(3)}R d^3x$ , etc., which can be constructed out of products of the Riemann tensor and its covariant derivatives, with the topology of 3-space itself being separately specified.

Higgs<sup>8</sup> has pointed out that in an infinite world such a characterization of  $\Psi$  would be inadequate, for in this case the asymptotic coordinates also play a role.  $\Psi$  could instead be represented as a functional of any three of the six coordinate-invariant functions<sup>22</sup>:

$$\varphi^{AB}(\eta) \equiv \int \gamma^{1/2} \gamma^{ij} \zeta^A{}_{,i} \zeta^B{}_{,j} \delta^3(\eta - \zeta(x)) d^3x, \quad A, B = 1, 2, 3, \quad (5.3)$$

where the  $\zeta$ 's are scalars satisfying the elliptic differential equation

$$\zeta^A{}_{,i}{}^i = 0, \quad (5.4)$$

with the boundary conditions  $\zeta^A \rightarrow x^A$  at infinity. The  $\zeta$ 's define a harmonic coordinate system, and Eqs. (5.4) yield  $\partial\varphi^{AB}/\partial\eta^B = 0$  as a corollary. If  $\varphi^{11}, \varphi^{12}, \varphi^{22}$  are arbitrarily chosen then  $\varphi^{13}, \varphi^{23}, \varphi^{33}$  are determined by integrating successively the equations  $\partial\varphi^{13}/\partial\eta^3 = -\partial\varphi^{11}/\partial\eta^1 - \partial\varphi^{12}/\partial\eta^2, \partial\varphi^{23}/\partial\eta^3 = -\partial\varphi^{12}/\partial\eta^1 - \partial\varphi^{22}/\partial\eta^2, \partial\varphi^{33}/\partial\eta^3 = -\partial\varphi^{13}/\partial\eta^1 - \partial\varphi^{23}/\partial\eta^2$ . If space-time is asymptotically flat and the coordinates  $x^i$  are Minkowskian at infinity, then these equations can be consistently integrated with the asymptotic boundary conditions  $\varphi^{AB} \rightarrow \delta_{AB}$ .

The above example is cited in order to re-emphasize the fundamental difference between finite and infinite worlds. In the finite case we may replace the symbol  $\Psi[\gamma]$  by  $\Psi[{}^{(3)}G]$  to display the fact that  $\Psi$  depends only on the 3-geometry, denoted here by  ${}^{(3)}G$ , and on nothing else, whereas in the infinite case we must write something like  $\Psi[{}^{(3)}G, \mathcal{L}]$ , with  $\mathcal{L}$  symbolizing the surrounding laboratory which determines the asymptotic coordinates (including, in the Schrödinger picture, the coordinate  $x^0$ ).

We shall denote by  $\mathfrak{M}$  the set of all possible 3-geometries which a finite world may possess. The following question will arise: Can a topology be imposed upon  $\mathfrak{M}$  which is both meaningful and at the same time useful in the context of the quantum theory of finite worlds? One possibility which suggests itself is to view  $\mathfrak{M}$  as an infinite-dimensional vector space whose "points" are discrete sets of invariants mentioned earlier. The topology could be that defined by the Cartesian metric on this space, and the symbol  ${}^{(3)}G$  could be replaced by a set of vector components. In fact, this possibility is not very useful, and although we shall actively pursue the question of assigning a metric, and indeed a pseudo-Riemannian structure, to  $\mathfrak{M}$ , no advantage will be gained by attempting to make our symbolism more explicit. It will be sufficient simply to keep in mind the idea that  $\mathfrak{M}$  is not just a mere set but is actually a

<sup>22</sup> In the Schrödinger picture  $\Psi$  would also depend on  $x^0$ .

manifold. Thus we shall say:  $\mathcal{M}$  is the *domain manifold* for the state functional  $\Psi$ , and the  ${}^{(3)}G$  are its “points.”

So far nothing has been said about dynamics. The only way in which dynamics can enter the picture is through the Hamiltonian constraint. This now takes the form

$$\left( G_{ijkl} \frac{\delta}{\delta \gamma_{ij}} \frac{\delta}{\delta \gamma_{kl}} + \gamma^{1/2} {}^{(3)}R \right) \Psi[{}^{(3)}G] = 0, \quad (5.5)$$

where

$$G_{ijkl} \equiv \frac{1}{2} \gamma^{-1/2} (\gamma_{ik}\gamma_{jl} + \gamma_{il}\gamma_{jk} - \gamma_{ij}\gamma_{kl}). \quad (5.6)$$

According to our rule of freely commuting field operators taken at the same space-time point, the functional differential operator  $\delta/\delta \gamma_{ij}$  must always be understood to give zero when acting on a  $\gamma_{kl}$  at the same point.<sup>23</sup> If it were not for this rule, we might try to regard the first term in the parentheses of (5.5) as a kind of Laplace-Beltrami operator in a 6-dimensional Riemannian manifold having  $G_{ijkl}$  as its *contravariant* metric. Although such an interpretation is inappropriate for the operator itself, it is nevertheless useful to regard  $G_{ijkl}$  as a metric tensor and to study the properties of the manifold which it defines. These properties, which are derived in Appendix A, turn out to be quite interesting.

The manifold in question will be denoted by  $M$ . When  $\gamma_{ij}$  is positive definite (as it is for a spacelike hypersurface)  $M$  has the hyperbolic signature  $-+++++$ . A “pure dilation” of  $\gamma_{ij}$  (i.e., multiplication by a multiple of the unit matrix) constitutes a typical “timelike” displacement. It is convenient to introduce the timelike coordinate

$$\xi \equiv (32/3)^{1/2} \gamma^{1/4} \quad (5.7)$$

and any five other coordinates  $\xi^A$  orthogonal to it. The covariant metric then takes the form

$$\begin{pmatrix} -1 & 0 \\ 0 & (3/32)\xi^2 \bar{G}_{AB} \end{pmatrix}, \quad (5.8)$$

where

$$\bar{G}_{AB} \equiv \text{tr}(\gamma^{-1} \gamma_A \gamma^{-1} \gamma_B), \quad (5.9)$$

$$\gamma \equiv (\gamma_{ij}). \quad (5.10)$$

Expression (5.8) reveals  $M$  as a set of “nested” 5-dimensional submanifolds, all having the same intrinsic shape and differing only in the scale factor  $(3/32)\xi^2$ . The shape is described by the positive-definite metric  $\bar{G}_{AB}$  which, since expression (5.9) remains invariant under a dilation of the  $\gamma$ 's, is independent of  $\xi$ .

The manifold having  $\bar{G}_{AB}$  as a metric will be denoted by  $\bar{M}$ . It is shown in Appendix A that the geodesic

<sup>23</sup> There is nothing automatically pathological, however, about having two functional derivatives acting at the same point, as in (5.5). For example, if  $I \equiv \frac{1}{2} \int dx \int dx' \varphi(x) K(x, x') \varphi(x')$ , where  $\varphi$  is an arbitrary function and  $K$  is a fixed kernel, then  $\delta^2 I / \delta \varphi(x) \delta \varphi(x') = K(x, x')$ . Pathology occurs only if  $K(x, x')$  is singular at  $x' = x$ .

equation in  $\bar{M}$  takes the form

$$\frac{d^2 \gamma}{ds^2} - \frac{d\gamma}{ds} \gamma^{-1} \frac{d\gamma}{ds} = 0, \quad \text{tr} \left( \gamma^{-1} \frac{d\gamma}{ds} \right) = 0. \quad (5.11)$$

This has the general solution<sup>24</sup>

$$\gamma(s) = \mathbf{M} \sim e^{\mathbf{N}s} \mathbf{M}, \quad (5.12)$$

where  $\mathbf{M}$  is an arbitrary nonsingular  $3 \times 3$  matrix and  $\mathbf{N}$  is subject only to the restrictions

$$\mathbf{N} \sim \mathbf{N}, \quad \text{tr} \mathbf{N} = 0, \quad \text{tr} \mathbf{N}^2 = 1, \quad (5.13)$$

the last of which guarantees that  $s$  is the arc length. Since  $e^{\mathbf{N}s}$  is analytic for all values of  $s$ ,  $\bar{M}$  is geodesically complete. It is not difficult to show that any two points of  $\bar{M}$  may be joined by a unique geodesic and that if the two points are represented by symmetric matrices  $\gamma_1$  and  $\gamma_2$  having the same determinant then their distance of separation is  $\{\text{tr}[\ln(\gamma_1^{-1}\gamma_2)]^2\}^{1/2}$ . The manifold  $\bar{M}$  is evidently noncompact and diffeomorphic to Euclidean 5-space.

By straightforward computation one may verify that the Riemann and Ricci tensors of  $\bar{M}$  have the respective forms

$$\bar{R}_{ABCD} = \text{tr} [\gamma^{-1} \gamma_{,D} \gamma^{-1} \gamma_{,C} \gamma^{-1} \times (\gamma_{,A} \gamma^{-1} \gamma_{,B} - \gamma_{,B} \gamma^{-1} \gamma_{,A})], \quad (5.14)$$

$$\bar{R}_{AB} = -\frac{3}{2} \bar{G}_{AB}. \quad (5.15)$$

From the latter it follows that  $\bar{M}$  is an “Einstein space” of constant negative Gaussian curvature. It is furthermore not difficult to show that the Riemann tensor (5.14) has vanishing covariant derivative, which implies that  $\bar{M}$  is, in fact, a *symmetric space*<sup>25</sup> with a certain group structure. The group structure may be deduced from the observation that the transformation

$$\gamma' = \mathbf{L} \sim \gamma \mathbf{L}, \quad (5.16)$$

where  $\mathbf{L}$  is an arbitrary constant nonsingular  $3 \times 3$  matrix, leaves the metric (5.9) unchanged. The full linear group in three dimensions therefore acts isometrically on  $\bar{M}$ . Because of the dilation invariance of the points of  $\bar{M}$ , however, it is only the simple Lie Group  $SL(3, R)$  which acts effectively on it. It is easily verified that  $SL(3, R)$  acts transitively on  $\bar{M}$  and, moreover, that the *isotropy subgroup*<sup>26</sup> at any point is isomorphic to  $SO(3)$ .  $\bar{M}$  may therefore be identified as the coset space

$$\bar{M} = SL(3, R) / SO(3). \quad (5.17)$$

Although the manifold  $\bar{M}$  is geodesically complete, the manifold  $M$  is not. It is shown in Appendix A that all geodesics in  $M$  ultimately hit a *frontier* of infinite

<sup>24</sup> The tilde “ $\sim$ ” denotes the transpose. All matrices are assumed real.

<sup>25</sup> See, for example, S. Helgason, *Differential Geometry and Symmetric Spaces* (Academic Press Inc., New York, 1962). The author is indebted to Professor Helgason for enlightenment as to the group structure of  $\bar{M}$ .

curvature. "Timelike" and null geodesics hit it at one end; "spacelike" geodesics hit it at both ends. This frontier, which will be denoted by  $F$ , is located at  $\zeta=0$ , as may be inferred from the readily computed curvature scalar

$${}^{(6)}R = -60/\zeta^2. \quad (5.18)$$

A question now arises as to what extent the Riemannian structure of  $M$  may be regarded as imposing a structure on  $\mathfrak{M}$  by way of the Hamiltonian constraint. Without attempting to answer this question directly, we may point out certain very suggestive features of the theory. First of all, the existence of the timelike coordinate  $\zeta$  in  $M$  suggests that a corresponding "intrinsic time" exists in  $\mathfrak{M}$  and that the Hamiltonian constraint does indeed have dynamical content. This idea is given support by the following considerations: The specification of a given 3-geometry requires the assignment of essentially 3 independent quantities at each point of 3-space. If we regard the usual enumeration of the degrees of freedom possessed by the gravitational field, namely *two* for every point of 3-space, as being valid in a finite world, this leaves one quantity per 3-space point to play the role of intrinsic time. Baierlein, Sharp, and Wheeler<sup>11,26</sup> have shown in the classical theory that if the intrinsic geometry is given on any two hypersurfaces then, except in certain singular cases, the geometry of the entire space-time manifold, *and hence the absolute time lapse between the two hypersurfaces*, is determined. Moreover, it is determined solely by the constraints. Analogously, the quantum theory is completely determined by the transformation functional  $\langle {}^{(3)}G' | {}^{(3)}G'' \rangle$ , where  $| {}^{(3)}G\rangle$  denotes that state of the gravitational field for which there exists at least one hypersurface having an infinitely precise geometry  ${}^{(3)}G$ . Wheeler<sup>11</sup> has emphasized the importance of the two-hypersurface formulation of gravodynamics (or "geometrodynamics" as he calls it) and has suggested the use of the Feynman sum-over-histories method to compute the transformation functional.<sup>27</sup>

Another suggestive feature of the theory is the following. Because of the hyperbolic character of  $M$  the Hamiltonian constraint (5.5) resembles a Klein-Gordon equation, with  $-\gamma^{1/2} {}^{(3)}R$  playing the role of the mass-squared term. An important difference, however, is that  ${}^{(3)}R$  can be either positive or negative, and hence the "wave" propagation of the state functional is not confined to timelike directions.

<sup>26</sup> R. F. Baierlein, D. H. Sharp, and J. A. Wheeler, Phys. Rev. **126**, 1864 (1962). There is nothing mysterious about the existence of a manifold of "time" variables. The same manifold exists in conventional field theory in those formulations which make the state functional depend on an arbitrary spacelike hypersurface.

<sup>27</sup> The sum-over-histories or "functional integral" method has not yet been applied to any "practical" problem of quantum gravodynamics. It will be encountered in heuristic and formal applications in the following papers of this series. Its consistency with the Dirac theory has been demonstrated by Leutwyler. [See H. Leutwyler, Phys. Rev. **134**, B1155 (1964).]

In spite of this difference the analogy with the Klein-Gordon theory suggests the following definition for the quantum-mechanical inner product of two states  $\Psi_a$  and  $\Psi_b$ :

$$\begin{aligned} (\Psi_b, \Psi_a) &= Z \int_{\Sigma} \Psi_b^* [{}^{(3)}G] \\ &\times \prod_x \left( d\Sigma^{ij} G_{ijkl} \frac{\delta}{i\delta\gamma_{kl}} - \frac{\delta}{i\delta\gamma_{kl}} G_{ijkl} d\Sigma^{ij} \right) \Psi_a [{}^{(3)}G]. \end{aligned} \quad (5.19)$$

The infinite product, which arises because (5.5) is really not just one equation but  $\infty^3$  equations, is here taken over all the points of 3-space, and is to be understood in a formal sense as representing the result of a limiting process based on a sequence of lattices in 3-space, each lattice requiring the introduction of a corresponding normalizing constant  $Z$ . The symbol  $\Sigma$  denotes the topological product of a set of 5-dimensional  $M$ -hypersurfaces  $\Sigma(x)$  (one chosen at each point of 3-space), the  $d\Sigma^{ij}$  being their directed surface elements. It is an immediate consequence of the Hamiltonian constraint that this inner product is independent of the choice of  $\Sigma(x)$ 's provided some kind of appropriate boundary conditions are satisfied at the "edges" of  $\Sigma$ . It is also worth noting that since the  $G_{ijkl}$  do not involve any spatially differentiated  $\gamma$ 's, the operators standing in the infinite product all commute, and hence no factor-ordering difficulties arise here.

In view of the coordinate invariance of the state functionals the inner product integral (5.19) contains a  $3 \times \infty^3$ -fold redundancy arising from the geometrical indistinguishability of 3-metrics which differ only by coordinate transformations.<sup>28</sup> This produces a divergence which must be formally absorbed into the normalization constant  $Z$ , and reminds us that  $\mathfrak{M}$  is not just the topological product of  $M$  with itself over all the points of 3-space, but is a subspace of the latter manifold.

Another difficulty with the definition (5.19) concerns the problem of "negative probability." This problem arises here, just as it does for the Klein-Gordon equation, from the fact that the Hamiltonian constraint involves a second derivative with respect to the "time" coordinate. If the  $\Sigma(x)$ 's are chosen "spacelike," then the only way to assure positive definiteness of (5.19), when  $\Psi_b = \Psi_a$ , is to restrict the content of  $\Psi_a$  to "positive frequency" components with respect to every "time" coordinate  $\zeta(x)$ . Restriction to such components, however, implies that  $\Psi_a$  vanishes nowhere in the range  $-\infty < \zeta < \infty$ , and this conflicts with the one-sided character of  $\zeta$ , namely  $\zeta > 0$ , which follows from the geometrical analysis revealing the existence of a frontier in  $M$  at  $\zeta = 0$ . One might hope that an analytic continuation could be performed around  $\zeta = 0$ , but

<sup>28</sup> A coordinate transformation generally produces a change in  $\Sigma$ , but this does not affect the integral.

whether this would have any physical meaning is unclear. The singularity in the Hamiltonian constraint at  $\xi=0$  is a strong one, as may be seen by rewriting (5.5) in the form

$$\left[ -\frac{\delta^2}{\delta\xi^2} + \frac{(32/3)\bar{G}^{AB}}{\xi^2} \frac{\delta^2}{\delta\xi^A \delta\xi^B} + (3/32)\xi^2 {}^{(3)}R \right] \times \Psi [{}^{(3)}g] = 0, \quad (5.20)$$

which makes use of (5.7) and (5.8). The question at issue is whether the frontier in  $M$  generates a corresponding barrier in  $\mathfrak{M}$  beyond which there is no possibility of extending the state functional. Unfortunately, in the present state of our knowledge no clear-cut answer can be given to this question. Some of the problems which have a bearing on it, however, can be identified. These will now be discussed.

## 6. THE METRIC OF $\mathfrak{M}$ . THE HAMILTON-JACOBI EQUATION AND GRAVITATIONAL COLLAPSE

The most obvious way to approach  $\mathfrak{M}$  is through the manifold  $M^{\infty^3}$ , which is defined formally as the topological product of  $M$  with itself over the points of 3-space:

$$M^{\infty^3} \equiv \prod_{\mathbf{x}} M(\mathbf{x}). \quad (6.1)$$

The “points” of  $M^{\infty^3}$  are the matrix functions  $\gamma_{ij}(\mathbf{x})$ . For brevity they will be denoted simply by  $\gamma$ . In practice the definition (6.1) must be supplemented by some sort of continuity requirements. For example,  $\gamma$  may be required to be continuous and piecewise differentiable. However, we do not wish to be precise about this here, since as yet no rigorous theory of the role of the manifold  $\mathfrak{M}$  in the quantum theory exists. We wish merely to point out some of the issues involved, and to leave the formalism itself as unencumbered as possible. Thus we shall be willing to admit any sort of pathology for  $\gamma$  which we can get away with, i.e., for which some sort of physical interpretation exists, however idealized, which permits  $\gamma$  to be handled in a consistent fashion. For example, geometrical singularities at which the Riemann tensor behaves like a differentiated  $\delta$  function, or for which integrals like  $\int \gamma^{1/2} {}^{(3)}R dx$  still exist, will not be excluded *a priori*. In the same spirit, we shall not place any restrictions on coordinate transformations beyond perhaps requiring them to be differentiable (so that tensor transformation laws exist almost everywhere) and one-to-one (so that they form a group). Thus we shall not automatically exclude transformations for which the Jacobian either vanishes or diverges at certain points. The ultimate question will always be: What is the *barrier* beyond which we cannot go? In every case this will probably depend, to some extent at least, on the context, and we do not wish to prejudice the answer in advance.

There is, however, one trivial pathology which may

be avoided without loss of generality, namely, coordinate singularities which arise from the impossibility of covering compact manifolds with a single well-behaved coordinate system. We shall always assume that 3-space is covered with a finite set of overlapping coordinate patches, each of which can be put into one-to-one correspondence with a certain portion of the Cartesian mesh in Euclidean 3-space, and on the boundaries of which the coordinates are held fixed. In addition a set of supplementary connection formulas between patches must be assumed to hold in the overlap regions. All of this paraphernalia is to be understood as included in the definition (6.1), which means that each function  $\gamma_{ij}(\mathbf{x})$  is really a set of functions, one in each coordinate patch, and that the  $\mathbf{x}$ 's in Eq. (6.1) are to be understood as ranging over all values in all patches.

Now let  $\gamma$  to be a fixed point of  $M^{\infty^3}$ . Consider the set of all points which may be reached from  $\gamma$  by coordinate transformations. This set is known as the *orbit* of  $\gamma$  under the coordinate transformation group and will be denoted by “orb  $\gamma$ .” There is a one-to-one correspondence between the orbits in  $M^{\infty^3}$  and the points of  $\mathfrak{M}$ . In fact no generality is lost if they are identified:

$$\text{orb } \gamma \equiv {}^{(3)}g. \quad (6.2)$$

Suppose  $M^{\infty^3}$  is endowed with a metric. (That this is feasible will appear in a moment.) If this metric satisfies a certain condition then it will impose, in a natural way, a metric on  $\mathfrak{M}$ . The condition is that the coordinate transformation group in 3-space be an *isometry group* of  $M^{\infty^3}$ . The associated metric in  $\mathfrak{M}$  is then obtained by defining the distance between two neighboring orbits to be the shortest distance in  $M^{\infty^3}$ .

It is shown in Appendix B that the above condition is satisfied if and only if the metric in  $M^{\infty^3}$  transforms, under 3-dimensional coordinate transformations, contragrediently to the Kronecker product  $\gamma_{ij}\gamma_{k'l'}$ . This means that the metric in  $M^{\infty^3}$ , which we shall denote by  $g^{ijk'l'}$ , must be a contravariant bitensor density of weight at both  $\mathbf{x}$  and  $\mathbf{x}'$ .

There are infinitely many contravariant bitensor densities which can be constructed out of the  $\gamma$ 's and which might serve as acceptable metrics for  $M^{\infty^3}$ . Of these, however, there is only a single one-parameter family which is *local*, i.e., which involves only undifferentiated  $\gamma$ 's and for which both  $g^{ijk'l'}$  and its inverse vanish when  $\mathbf{x} \neq \mathbf{x}'$ . This family is given by

$$g^{ijk'l'} = \frac{1}{2} \gamma^{1/2} (\gamma^{ik}\gamma^{jl} + \gamma^{il}\gamma^{jk} + \lambda \gamma^{ij}\gamma^{kl}) \delta(\mathbf{x}, \mathbf{x}'), \quad (6.3)$$

where  $\lambda$  can assume any real value except  $-\frac{2}{3}$ . If we wished to impose a positive-definite metric on  $\mathfrak{M}$ , so that we could use, as the condition for the identity of two 3-geometries, the vanishing of the “distance” between them, then the metric of  $M^{\infty^3}$  itself would have to be positive definite. In the present case this requires  $\lambda > -\frac{2}{3}$ , the simplest choice being  $\lambda=0$ . On the other

hand, the choice  $\lambda = -2$  is the natural choice if we assume that Eq. (6.1) defines not merely a topological product but also a geometrical structure generated by the original metric on  $M$ . For then we have

$$\int \mathcal{G}_{ij} a'' b' \mathcal{G}^{a'' b' k' l'} d^3 x'' = \delta_{ij} k' l', \quad (6.4)$$

where

$$\mathcal{G}_{ijkl} \equiv G_{ijkl} \delta(\mathbf{x}, \mathbf{x}'), \quad (6.5)$$

with  $G_{ijkl}$  given by Eq. (5.6). In this case the “arc length”  $d\mathfrak{s}$  associated with a displacement  $d\gamma_{ij}$  in  $M^{\infty 3}$  is given by

$$\begin{aligned} d\mathfrak{s}^2 &= \int d^3 x \int d^3 x' \mathcal{G}^{ijk'l'} d\gamma_{ij} d\gamma_{kl} \\ &= \int \gamma^{1/2} (\gamma^{ik} \gamma^{jl} - \gamma^{ij} \gamma^{kl}) d\gamma_{ij} d\gamma_{kl} d^3 x. \end{aligned} \quad (6.6)$$

Geodesics in  $M^{\infty 3}$  are not, in general, geodesics in  $\mathfrak{M}$ . However, in Appendix B it is shown that if a geodesic in  $M^{\infty 3}$  intersects one of the orbits in its path orthogonally then it is a geodesic in  $\mathfrak{M}$ , and, moreover, it intersects every other orbit in its path orthogonally. This means that it is in principle possible to use formula (A69) of the Appendix to determine the distance between two 3-geometries. In practice, of course, the amount of labor involved is formidable, assuming that the 3-geometries are given in the form of two matrix functions  $\gamma_1(\mathbf{x})$  and  $\gamma_2(\mathbf{x})$ . One must integrate expression (A69) over 3-space and then find the minimum of the integral as one of the functions, say  $\gamma_1(\mathbf{x})$ , is held fixed while the other ranges over the various equivalent forms it can take under coordinate transformations. This means solving the complicated set of nonlinear partial differential equations which result from the corresponding variational principle and which, in effect, yield the coordinate transformation which “lines up”  $\gamma_1(\mathbf{x})$  and  $\gamma_2(\mathbf{x})$ , so that a geodesic from one to the other intersects orbits orthogonally.

Such complications can be avoided if one merely wants to know the distance from a given 3-geometry to the frontier.<sup>29</sup> In this case, since the frontier is an extended object and is at different distances—spacelike, timelike, and null—in different directions, it is necessary to specify a direction  $d\gamma_{ij}/d\mathfrak{s}$  in addition to the 3-

<sup>29</sup> By *frontier* we do not necessarily mean *barrier*. It must be repeatedly emphasized that very little is known about the general conditions under which extensions beyond the frontier can be carried out. Here we are defining the frontier to be simply the locus of points  $\gamma$  in  $M^{\infty 3}$  for which the matrix  $\gamma_{ij}(\mathbf{x})$  has one or more singularity points ( $\gamma=0$ ) in 3-space, regardless of whether or not these singularity points represent real geometrical singularities. A formal definition would be

$$F = \bigcup_{\mathbf{x}} F(\mathbf{x}) \coprod_{\mathbf{x}' \neq \mathbf{x}} M(\mathbf{x}'),$$

where  $\coprod$  and  $\cup$  denote the topological product and union, respectively, and  $F(\mathbf{x})$  is the frontier of  $M(\mathbf{x})$ .

geometry itself. Here  $\mathfrak{s}$  is either the arc length (6.6) or, in the exceptional null case, an affine parameter, and  $d\gamma_{ij}/d\mathfrak{s}$  must satisfy the starting condition [cf. Eq. (B24)]

$$\gamma^{jk} \left[ \left( \frac{d\gamma_{ij}}{d\mathfrak{s}} \right)_{,k} - \left( \frac{d\gamma_{jk}}{d\mathfrak{s}} \right)_{,i} \right] = 0, \quad (6.7)$$

which guarantees that the starting direction will be orthogonal to the starting orbit. The square of the distance in the frontier in the assigned direction is then given by<sup>30</sup>

$$\mathfrak{s}^2 = \min 2\sigma(\gamma, d\gamma/d\mathfrak{s}) \left[ G^{ijkl} \frac{d\gamma_{ij}}{d\mathfrak{s}} \frac{d\gamma_{kl}}{d\mathfrak{s}} \right]^{-1}, \quad (6.8)$$

where  $G^{ijkl}$  and  $\sigma$  are defined in Appendix A, Eqs. (A1) and (A63), respectively, and “min” denotes the minimum value over 3-space. The metric tensor at the point (or points) in 3-space at which the minimum occurs will become singular when the “point”  $\gamma$  in  $M^{\infty 3}$  has progressed a distance  $\mathfrak{s}$  along the geodesic. The geodesic can then go no further without changing the signature of a portion of 3-space. The frontier has been reached.

It does not automatically follow that 3-space acquires a *geometrical* singularity at the frontier. However, there are several facts worth noting.

(1) The occurrence of a singular metric cannot be avoided by changing the coordinates as one proceeds along the geodesic. Although it is true that a coordinate transformation can carry one from one point to another in  $M^{\infty 3}$  and even, seemingly, away from the frontier, yet since expression (6.8) is a scalar,  $\mathfrak{s}$  remains unchanged. What happens is that the coordinate transformation also changes the direction  $d\gamma_{ij}/d\mathfrak{s}$ . Moreover, the covariance of Eq. (6.7) ensures that the orthogonality of the geodesic to the orbits is left unaffected.

(2) As long as no coordinate transformations are performed while  $\gamma$  is moving along its orthogonal geodesic, the coordinate system in 3-space, if initially nonsingular, will remain nonsingular until the frontier is reached. No such statement can be made for nonorthogonal geodesics, which in some cases follow a circuitous route in  $\mathfrak{M}$  from a given  ${}^{(3)}\mathcal{G}$  back again to the same  ${}^{(3)}\mathcal{G}$ , but in a different coordinate system.<sup>31</sup>

(3) It is not necessary that the metric become singular simultaneously at all points of 3-space in order that

<sup>30</sup> When  $d\gamma_{ij}/d\mathfrak{s}$  is a null vector in  $M$ , expression (6.8) becomes an indeterminate form 0/0. At such points in 3-space (6.8) may, in view of Eq. (A54) which implies  $\mathfrak{s}=\text{constant} \times \gamma^{1/2}$ , be replaced simply by  $\mathfrak{s}^2 = 4(\gamma^{ij} d\gamma_{ij}/d\mathfrak{s})^{-2}$ .

<sup>31</sup> In attempting to visualize  $M^{\infty 3}$  and  $\mathfrak{M}$  it is helpful to have a simpler model in mind. The following is suggested: Let the big manifold be Euclidean 3-space and let the group be rotations about an axis. The orbits are then circles concentric with the axis and at right angles to it, and the orbit manifold is the Euclidean half-plane. A straight line (i.e., geodesic) in the big manifold will be a geodesic in the orbit manifold if it intersects or is parallel to the axis, so that it intersects every circle in its path at right angles. A straight line which is skew to the axis, however, is a hyperbola in the orbit manifold, and a skew line at right angles to the axis returns again to each orbit which it intersects.

the frontier be reached. On the other hand, this *can* happen. It happens, for example, when  $d\gamma_{ij}/d\delta = \text{constant} \times \gamma_{ij}$ , which obviously satisfies (6.7). In this case the geodesic motion is one of pure dilation, and the square of the distance to the frontier is, apart from an unimportant factor, simply the volume of 3-space.

(4) In the case of a pure dilation it is obvious that a geometrical singularity (zero volume) *does* occur at the frontier. That geometrical singularities must also occur in many other cases as well follows from the readily verified relation

$$d^{(3)}R/d\delta = -^{(3)}R^{ij}d\gamma_{ij}/d\delta, \quad (6.9)$$

which holds as long as condition (6.7) is satisfied. In Appendix A it is shown that most geodesics (i.e., all but a set of measure zero) strike the frontier at points where some of the  $\gamma_{ij}$  (and hence some of the  $d\gamma_{ij}/d\delta$ ) become infinite, even though  $\gamma$  itself vanishes. Except in special cases, therefore, expression (6.9) will acquire singularities at the frontier.

With these mathematical preliminaries in mind let us now have a look at quantum dynamics. It is helpful to begin by analyzing Eq. (5.5) in the WKB approximation, so as to make the maximum possible use of classical ideas. We write

$$\Psi^{(3)}[\mathcal{G}] = \mathcal{A} \exp(i\mathcal{W}), \quad (6.10)$$

where  $\mathcal{A}$  and  $\mathcal{W}$  are assumed to be real functionals satisfying (roughly) the restriction

$$|\delta\mathcal{A}/\delta\gamma_{ij}| \ll |\mathcal{A}\delta\mathcal{W}/\delta\gamma_{ij}|. \quad (6.11)$$

The phase then satisfies the Hamilton-Jacobi equation<sup>32</sup>

$$G_{ijkl} \frac{\delta\mathcal{W}}{\delta\gamma_{ij}} \frac{\delta\mathcal{W}}{\delta\gamma_{kl}} = \gamma^{1/2} {}^{(3)}R, \quad (6.12)$$

while the amplitude satisfies the conservation law

$$\delta(\mathcal{A}^2 G_{ijkl} \delta\mathcal{W}/\delta\gamma_{kl})/\delta\gamma_{ij} = 0. \quad (6.13)$$

In addition, the  $\chi$  constraints impose the restrictions

$$\left( \frac{\delta\mathcal{W}}{\delta\gamma_{ij}} \right)_{,j} = 0, \quad \left( \frac{\delta\mathcal{A}}{\delta\gamma_{ij}} \right)_{,j} = 0. \quad (6.14)$$

Each solution of the Hamilton-Jacobi equation (6.12) determines a family of solutions of the classical field equations (i.e., a family of Ricci-flat 4-geometries) having the following property: For every 3-geometry there exists one and only one member of the family which has the 3-geometry as a spacelike hypersection, i.e., for which the 3-geometry is to be found among the infinity of spacelike hypersections which the member admits. Once  $\mathcal{W}$  is given, each 3-geometry determines a

<sup>32</sup> The Hamilton-Jacobi equation for general relativity appears to have been first written down by A. Peres, Nuovo Cimento **26**, 53 (1962).

unique 4-geometry. The 4-geometry may be computed by making the identification  $\pi^{ij} = \delta\mathcal{W}/\delta\gamma_{ij}$  and integrating the equation

$$\partial\gamma_{ij}/\partial x^0 = 2\alpha G_{ijkl} \delta\mathcal{W}/\delta\gamma_{kl} + \beta_{i,j} + \beta_{j,i}, \quad (6.15)$$

which follows from (2.5) and (3.3).<sup>33</sup> The quantities  $\alpha$  and  $\beta_i$  are, as always, completely arbitrary, at least in sufficiently small finite regions. (Some global restrictions will generally exist.)

It is not hard to verify that (6.15) does indeed yield a solution of the classical field equations for each initial 3-geometry. One simply differentiates (6.15) with respect to  $x^0$  and replaces  $\delta\mathcal{W}/\delta\gamma_{kl}$  by its expression in terms of  $\alpha$ ,  $\beta_{i,j}$ , and  $\partial\gamma_{ij}/\partial x^0$ . One finds

$$\begin{aligned} \gamma_{ij,00} &= (\ln\alpha)_{,0}(\gamma_{ij,0} - \beta_{i,j} - \beta_{j,i}) + \frac{\partial G_{ijkl}}{\partial\gamma_{mn}} \\ &\quad \times G^{klrs} \gamma_{mn,0}(\gamma_{rs,0} - 2\beta_{r,s}) + 4\alpha G_{ijkl} \int \frac{\delta^2}{\delta\gamma_{kl} \delta\gamma_{m'n'}} \\ &\quad \times \left( \alpha' G_{m'n'r's'} + \beta_{m'n'} \right) d^3x'. \end{aligned} \quad (6.16)$$

The integration which appears in the last term of this equation may be performed with the aid of the identities

$$\begin{aligned} G_{m'n'r's'} &= \frac{\delta^2\mathcal{W}}{\delta\gamma_{kl} \delta\gamma_{m'n'}} \frac{\delta\mathcal{W}}{\delta\gamma_{r's'}} \\ &= -\frac{1}{2} \left[ \frac{\partial G_{mnrs}}{\partial\gamma_{kl}} \frac{\delta\mathcal{W}}{\delta\gamma_{mn}} \frac{\delta\mathcal{W}}{\delta\gamma_{rs}} + \gamma^{1/2} ({}^{(3)}R^{kl} - \gamma^{kl} {}^{(3)}R) \right] \\ &\quad \times \delta(x, x') + \frac{1}{2} \gamma'^{1/2} \gamma^{m'n'} \gamma^{r's'} \\ &\quad \times (\delta_{m'n'r's'} - \delta_{m'n'r's'}), \end{aligned} \quad (6.17)$$

$$\left( \frac{\delta^2\mathcal{W}}{\delta\gamma_{kl} \delta\gamma_{m'n'}} \right)_{,n'} = \frac{\delta\mathcal{W}}{\delta\gamma_{n'r'}} \left( \frac{1}{2} \delta_{n'r'}^{kl} m' - \delta_{n'r'}^{kl} r' \right), \quad (6.18)$$

which are obtained by functionally differentiating Eqs. (6.12) and (6.14) and making use of (4.25). The result is a set of six local-field equations which, together with (6.12) and (6.14) re-expressed in terms of  $\alpha$ ,  $\beta_{i,j}$ ,  $\gamma_{ij,0}$ , are equivalent to the ten Einstein empty-space equations.

Let us now make the simplifying assumptions  $\alpha_{,i} = 0$ ,  $\beta_i = 0$ . We then have

$$G^{ijkl} \gamma_{ij,0} \gamma_{kl,0} = 4\alpha^2 \gamma^{1/2} {}^{(3)}R. \quad (6.19)$$

Let us also assume that the integral

$$I \equiv \int \gamma^{1/2} {}^{(3)}R d^3x \quad (6.20)$$

<sup>33</sup> The inverse problem of constructing the  $\mathcal{W}$  which corresponds to a given family of solutions of the classical field equations has been analyzed in detail by U. H. Gerlach (to be published). The author is indebted to Gerlach for the opportunity of studying this analysis in manuscript prior to publication.

(extended over the whole of 3-space) is nonvanishing. We may then choose

$$\alpha = \frac{1}{2} |I|^{-1/2}, \quad (6.21)$$

which permits  $x^0$  to be identified with the arc length  $\mathfrak{s}$  in the manifold  $M^{x^0}$  and permits the first of Eqs. (6.14) to be re-expressed in the form

$$(G^{ijkl} d\gamma_{kl}/d\mathfrak{s})_{,j} = 0, \quad (6.22)$$

which is identical with (6.7), showing that  $x^0$  is in fact also the arc length in  $\mathfrak{M}$ . Finally, Eq. (6.16) takes the form

$$\begin{aligned} \frac{d^2\gamma_{ij}}{d\mathfrak{s}^2} - \frac{d\gamma_{ik}}{d\mathfrak{s}} \frac{d\gamma_{ij}}{d\mathfrak{s}} + \frac{1}{2} \frac{d\gamma_{ij}}{d\mathfrak{s}} \frac{d\gamma_{kl}}{d\mathfrak{s}} + \frac{1}{8} \gamma^{-1/2} \gamma_{ij} G^{mnr} \frac{d\gamma_{mn}}{d\mathfrak{s}} \\ \times \frac{d\gamma_{rs}}{d\mathfrak{s}} = \frac{d \ln \alpha}{d\mathfrak{s}} \frac{d\gamma_{ij}}{d\mathfrak{s}} - 2\alpha^2 ({}^3R_{ij} - \frac{1}{4}\gamma_{ij} {}^3R). \end{aligned} \quad (6.23)$$

If the right-hand side of Eq. (6.23) were zero, the sequence of 3-geometries (as  $\mathfrak{s}$  varies) would trace out a geodesic in  $\mathfrak{M}$ . The right-hand term may therefore be regarded as a “force” term which caused the actual “trajectory” of 3-space to deviate from a geodesic. The following important questions arise: Is this “force term” powerful enough to keep the trajectory from striking the frontier? If not, what does arrival at the frontier mean physically?

Before giving answers to these questions, let us first take a crude over-all look at some of the simple implications of Eq. (6.23). It is not difficult to verify that if this equation is multiplied by  $\frac{1}{2}\gamma^{1/2}\gamma^{ij}$  and the result is integrated over 3-space, the following equation is obtained:

$$\frac{d^2V}{d\mathfrak{s}^2} - \frac{d \ln \alpha}{d\mathfrak{s}} \frac{dV}{d\mathfrak{s}} = -\frac{1}{4}, \quad V \equiv \int \gamma^{1/2} d^3x, \quad (6.24)$$

or equivalently,

$$\frac{d^2V}{d\tau^2} = -I, \quad (6.25)$$

where  $\tau$  is the “proper time”:

$$d\tau/d\mathfrak{s} = \alpha. \quad (6.26)$$

From this it follows that the curve of  $V$  as a function of  $\tau$  is concave downward whenever  $I$  is positive. Under these circumstances an expanding world tends to “slow down” while a contracting world tends to accelerate towards collapse.

A case for which  $I$  is positive is that in which 3-space has the geometry of a 3-sphere. The geometry cannot, however, remain spherical more than instantaneously, since the right-hand side of Eq. (6.19) is then everywhere positive, which requires the vector  $d\gamma_{ij}/d\mathfrak{s}$  to be “spacelike” in  $M$  for all  $\mathbf{x}$ , thus ruling out the possibility of a pure dilation. The derivative  $d\gamma_{ij}/d\mathfrak{s}$  must contain

shearing components corresponding to the presence of the gravitational radiation which is, in fact, needed in order to “close up” the universe. On the other hand, the 3-geometry may still be spherical in a coarse-grained sense. That is, although the sign of  $\gamma^{1/2} {}^3R$  may fluctuate at a fine-grained level due to the presence of gravitational waves, its mean value may approximate that of a 3-sphere. In this case Eq. (6.25) takes the approximate form

$$\frac{d^2V}{d\tau^2} \approx -6(4\pi^4 V)^{1/3}, \quad (6.27)$$

leading to a total lifetime of the universe given by<sup>34</sup>

$$\begin{aligned} T &\approx \frac{2}{3}(2\pi^2)^{-1/3} \int_0^{V_{\max}} \frac{dV}{(V_{\max}^{4/3} - V^{4/3})^{1/2}} \\ &= \sqrt{2} \operatorname{cn}^{-1}(0|\tfrac{1}{2}) R_{\max} = 2.62 R_{\max}, \end{aligned} \quad (6.28)$$

where  $R_{\max}$  is the radius of maximum expansion. Here it is clear that the “force term” in Eq. (6.23) does not prevent the 3-geometry from striking the frontier.

In the general case there are two factors which govern the trajectory of 3-space. Firstly, the condition  $\alpha_{,i}=0$ , which has been adopted in the above discussion, is known to be a poor one for keeping the hypersurfaces,  $x^0=\text{constant}$ , smooth. When these hypersurfaces are sandwiched together, as here, with spatially uniform intervals, they often quickly develop geometrical singularities which have nothing to do with the geometry of space-time.<sup>35</sup> Such singularities can usually be avoided simply by relaxing the condition  $\alpha_{,i}=0$ . However—and this is the second factor—it is now known from the work of Avez,<sup>36</sup> Penrose,<sup>37</sup> Hawking,<sup>38</sup> and Geroch<sup>39</sup> that a nontrivial singularity in space-time “almost always” occurs at some point in the history of any physically interesting universe. At such a point abandonment of the condition  $\alpha_{,i}=0$  is of no use. 3-space will acquire a geometrical singularity anyway. Thus, if the initial hypersurface is sufficiently close to the point of onset of a change in 3-space topology, or if a so-called “trapped 2-surface”<sup>37</sup> is on the point of being born within it, then it will develop a geometrical

<sup>34</sup> This is to be compared with  $T=2R_{\max}$  for a Friedmann universe filled with radiation treated as an ideal gas. Note that it is not possible to use expression (6.28) as an upper bound on the lifetime of the universe. Although it is easy to show that it is the spherical geometry which, for fixed  $V$ , makes  $I$  stationary (i.e., independent of small variations in the metric), this stationary point is neither a maximum nor a minimum, and hence it is not possible to assert that  $I \geq 6(4\pi^4 V)^{1/3}$ .

<sup>35</sup> The phenomenon occurs already in a flat space-time. It is not possible to construct a family of uniformly spaced curved spacelike hypersurfaces in Minkowski space without the members of the family developing a geometrical singularity either in the past or in the future. The singularity always develops in the convex direction, contrary to the situation in a Euclidean manifold.

<sup>36</sup> A. Avez, Ann. Inst. Fourier (Grenoble) **13**, 105 (1963).

<sup>37</sup> R. Penrose, Phys. Rev. Letters **14**, 57 (1965).

<sup>38</sup> S. W. Hawking, Phys. Rev. Letters **17**, 444 (1966).

<sup>39</sup> R. P. Geroch, Phys. Rev. Letters **17**, 445 (1966).

singularity which has nothing to do with the maintenance of the condition  $\alpha_{,i}=0$ . In this case the force term in (6.32) is again powerless to prevent the trajectory of 3-space from striking the frontier; indeed it may hasten the impact.

The occurrence of a singularity in space-time itself is known as *gravitational collapse*. Gravitational collapse may involve the whole of 3-space, as when the volume of the universe goes to zero, or it may involve only a small part of it (e.g., a collapsing superstar). It seems extremely likely that the almost universal inevitability of gravitational collapse is closely connected to the existence of the frontier in  $\mathfrak{M}$ . However, the establishment of this connection in rigorous terms is a major problem which remains unsolved. The existence proofs of Refs. 36–39 give no indication of the precise physical nature of the collapse singularity except for the statement that the normal causal properties of space-time break down there. This alone, of course, is enough to guarantee that the singularity represents a real barrier beyond which it is impossible to extend the solution of Einstein's equations. It means that in certain regions of the universe (or in the universe as a whole) time for the classical physicist, ultimately comes to an end beyond which he can make no further predictions.

The question now arises whether the classical collapse barrier, which we shall denote by  $\mathcal{B}$ , is also a barrier for the solutions of the quantum equation (5.5). That the answer is not obvious may be seen as follows. Consider first a point  $(^3)\mathcal{G}$  in  $\mathfrak{M}$  which is not on  $\mathcal{B}$ . If  $(^3)\mathcal{G}$  has a singularity this must be due to the hypersurface  $x^0=\text{constant}$  being chosen poorly. At the singular point in 3-space both the right and left sides of Eq. (6.12) will diverge. Correspondingly the two terms inside the parentheses of Eq. (5.5) will each contribute a divergence at this point. The two divergences will, however, cancel so that Eq. (5.5) is still satisfied. Consider now a point on  $\mathcal{B}$ . Here something special happens which causes Eq. (6.12) to break down. However, it does not automatically follow that Eq. (5.5) likewise breaks down, for there exist possibilities for treating Eq. (5.5) which have no counterparts in the classical theory. For example, we note that if  $\Psi$  is a solution of Eq. (6.12) then so is  $-\Psi$ . Moreover, the addition of an arbitrary constant to  $\Psi$  leaves Eq. (6.12) unaffected. Let this constant be adjusted so that  $\Psi$  vanishes at the point on  $\mathcal{B}$  in question, and choose for the WKB form of the solution of (5.5), the *superposition*

$$\Psi = \alpha [\exp(i\Psi) - \exp(-i\Psi)]. \quad (6.29)$$

Then  $\Psi$  itself vanishes at the barrier, and this might conceivably alleviate the singularity in Eq. (5.5) which would otherwise occur, and permit an extension of  $\Psi$  beyond the barrier.

If one is looking for an example on which to practice hand-waving arguments he might consider a situation in which 3-space is about to undergo a change in

topology. It can be shown that a change of topology requires (a) the development of a geometrical singularity in 3-space and (b) a breakdown in the causal structure (e.g., hyperbolic signature) of space-time at the onset of the singularity. Therefore topological transitions cannot be handled classically. However, since the singularity in  $(^3)\mathcal{G}$  need occur at only a single point of 3-space it may develop in such a way that the corresponding singularities of Eq. (5.5) all cancel. We are careful, of course, not to say that the singularities *will* cancel. No one really knows whether topological transitions can be handled quantum mechanically.

Although the classical and quantum barriers may not be identical, and although each may depend to some extent on the particular solution of the Hamilton-Jacobi equation (6.12), or of the "wave equation" (5.5), under consideration at the moment, it seems very probable that there exists an irreducible *core* which is common to all barriers. We have suggested that it may be possible to continue  $\Psi$  past 3-geometries which contain isolated singularities. However, it is extremely difficult to imagine how such a continuation could be performed beyond a 3-geometry which has a dense set of singularities, or which is singular at *all* of its points, e.g., a 3-space of zero volume. It is therefore likely that the following set theoretical inequality holds:

$$\text{orb} \prod_x F(x) \subseteq \mathcal{B}_Q \subseteq \mathcal{B}, \quad (6.30)$$

where  $\mathcal{B}_Q$  denotes the quantum barrier. In the remainder of the paper we shall assume that this inequality does hold.

The fact that  $\mathcal{B}$  is not the empty set is an embarrassment to the classical physicist, for it means that his theory breaks down. The fact that  $\mathcal{B}_Q$  is not the empty set, however, is not necessarily embarrassing to the quantum physicist, for he may be able to dispose of it by simply imposing, on the state functional, the following condition:

$$\Psi[(^3)\mathcal{G}] = 0 \quad \text{for all } (^3)\mathcal{G} \text{ on } \mathcal{B}_Q. \quad (6.31)$$

*Provided it does not turn out to be ultimately inconsistent*, this condition, which is already suggested by (6.29), yields two important results. Firstly, it makes the probability amplitude for catastrophic 3-geometries vanish, and hence gets the physicist out of his classical collapse predicament. Secondly, it may permit the Cauchy problem for the "wave equation" (5.5) to be handled in a manner very similar to that of the ordinary Schrödinger equation. Thus let  $\Sigma$  be a hypersurface like that which appears in Eq. (5.19). Since the dimensionality of  $\prod_x F(x)$  (the orbit of which forms the "core" of  $\mathcal{B}_Q$ ) is the same as that of  $\Sigma$ , namely  $5 \times \infty^3$ , it would appear that the specification of  $\Psi$  on  $\Sigma$ , together with the boundary condition (6.31), is equivalent to its specification on two hypersurfaces and hence suffices to determine  $\Psi[(^3)\mathcal{G}]$  completely for all  $(^3)\mathcal{G}$ .

If this heuristic argument [based on the analogy of Eq. (5.5) to the Klein-Gordon equation] is indeed valid, then it is not necessary to specify also the normal derivatives of  $\Psi$  on  $\Sigma$ , despite the fact that Eq. (5.5) is of the second differential order.<sup>40</sup>

The only obvious difficulty with condition (6.31) is that it makes the presence of “negative frequency” components in  $\Psi$  unavoidable (see the discussion at the end of Sec. 5) and hence leaves one very unclear as to how to use Eq. (5.19) to define inner products and at the same time maintain positive definiteness of probability. In the following sections we shall show how this difficulty can be resolved in a special case.

## 7. THE QUANTIZED FRIEDMANN UNIVERSE

The simplest classical model which exhibits the collapse phenomenon is the Friedmann universe. If the Friedmann universe is assumed to be closed it must be filled either with gravitational radiation or with some other form of energy. It is not difficult to show that when other forms of energy are present in addition to gravity, the Hamiltonian constraint condition (4.3) is replaced by

$$(\mathcal{H} + \mathcal{C})\Psi = 0, \quad (7.1)$$

where  $\mathcal{C}$  is the Hamiltonian of the system (or systems) giving rise to the additional energy. In order to avoid having to deal with entities as complicated as gravitons, with their spin and orbital states and their mutual interactions, we shall make use of such additional energy in the form of noninteracting material particles “at rest”. The Friedmann universe is obtained by distributing these particles uniformly throughout a 3-sphere and “freezing out” all the degrees of freedom of the gravitational field save one, namely, that which corresponds to the time-varying spherical radius  $R$ .

If  $\gamma_{ij}$  denotes the metric (in some coordinate system) of a 3-sphere of unit radius, then the 4-metric of the Friedmann world may be written in the form

$$(g_{\mu\nu}) = \begin{pmatrix} -\alpha^2 & 0 \\ 0 & \gamma_{ij} \end{pmatrix}, \quad \gamma_{ij} = R^2 \gamma^0_{ij}, \quad (7.2)$$

where  $\alpha$  and  $R$  depend only on  $x^0$ . Substituting this into (2.6), integrating over the volume  $2\pi^2 R^3$  of the Friedmann universe, and remembering that  $(^3)R$  for a 3-sphere is  $6/R^2$ , we obtain for the effective Lagrangian of the gravitational field

$$L = 12\pi^2 [-\alpha^{-1}R(R_{,0})^2 + \alpha R]. \quad (7.3)$$

As for the material particles (dust) which fill the universe, we shall, for reasons which will become clear as the analysis proceeds, endow them with internal dynamical degrees of freedom which may be described

<sup>40</sup> Alternative and more detailed heuristic arguments leading to the same conclusion have been given by H. Leutwyler, University of Bern report, 1965 (unpublished).

by canonical coordinates  $q^i$  and Lagrangians of the form  $l(q, \dot{q})$ , the dot denoting differentiation with respect to *proper* time:

$$\dot{q}^i = \alpha^{-1} q_{,0}^i. \quad (7.4)$$

Just as we have done for the gravitational field, however, we shall “freeze out” all the internal degrees of freedom save a small number by requiring all the particles to be identical and to be in coherent identical states (i.e., “in step”). Under these conditions the effective particle Lagrangian becomes

$$L = \alpha N l(q, \alpha^{-1} q_{,0}), \quad (7.5)$$

where  $N$  is the total number of the particles in the universe.

Adding (7.3) and (7.5) to obtain the total Lagrangian we see that once again we have the primary constraint

$$\pi = \partial(L + L)/\partial\alpha_{,0} = 0. \quad (7.6)$$

The wave function of the quantized Friedmann universe therefore cannot depend on  $\alpha$ .

The total Hamiltonian becomes

$$H + \mathbf{H} = \pi\alpha_{,0} + II R_{,0} + P_i q_{,0}^i - L - \mathbf{L} = \pi\alpha_{,0} + \alpha(\mathcal{H} + \mathcal{C}), \quad (7.7)$$

where

$$II = \partial L / \partial R_{,0} = -24\pi^2 \alpha^{-1} R R_{,0}, \quad (7.8)$$

$$P_i = \partial L / \partial \dot{q}^i = N p_i, \quad p_i = \partial l / \partial \dot{q}^i, \quad (7.9)$$

$$\mathcal{H} \equiv -II^2 / 48\pi^2 R - 12\pi^2 R, \quad (7.10)$$

$$\mathcal{C} \equiv Nm, \quad m \equiv p_i \dot{q}^i - l. \quad (7.11)$$

The symbol  $m$  is here used to denote the internal Hamiltonian of the particles because the Hamiltonian is, in fact, the rest mass, provided the arbitrary zero point of the Lagrangian  $l$  has been properly chosen. We note that the “kinetic energy” term in the gravitational Hamiltonian (7.10) has the opposite sign (i.e., negative) from that of conventional Hamiltonians. This is because the only motion permitted to a Friedmann universe is one of pure dilation, and hence the coordinate  $R$  is “timelike.”

The condition  $\pi_{,0} = 0$  leads immediately to the dynamical constraint  $\mathcal{H} + \mathcal{C} = 0$  which, in the quantum theory, takes the form (7.1). In the  $R$  representation this becomes

$$\left( \frac{1}{48\pi^2} R^{-1/4} \frac{\partial}{\partial R} R^{-1/2} \frac{\partial}{\partial R} R^{-1/4} - 12\pi^2 R + Nm \right) \Psi = 0, \quad (7.12)$$

where factors have been ordered in such a way that the first term inside the parentheses becomes a one-dimensional Laplace-Beltrami operator,<sup>41</sup> and  $m$  is now the particle mass operator. Equation (7.12), which is

<sup>41</sup> Here a definite ordering must be chosen. Since the number of degrees of freedom is now finite the ordering question cannot be treated as a problem in interpreting formally divergent symbols.

the analog of (5.5), must account completely for all the physical properties of the Friedmann universe.

It is by no means obvious how the familiar properties of the Friedmann world are to be extracted, in the classical limit, from Eq. (7.12), nor is it obvious what significance is to be attached to  $\Psi$  in the purely quantum domain. Difficulties of this type are not new to physics. A similar problem faced Schrödinger when he first wrote down the equation of the hydrogen atom. In his case there was a period of intense discussion, largely guided by Bohr, which ultimately led most physicists, with only a few dissenters of whom Einstein was the champion, to accept what has come to be known as the "Copenhagen view." The Copenhagen view depends on the assumed *a priori* existence of a classical level to which all questions of observation may ultimately be referred. Here, however, the whole universe is the object of inspection; there is no classical vantage point, and hence the interpretation question must be re-argued from the beginning. While we do not wish to stress this point unduly, since, after all, the Friedmann model ignores the vast complexities of the real universe, it is nevertheless clear that the quantum theory of space-time must ultimately force a deviation from the traditional Copenhagen doctrine.

Leaving aside these questions for the moment, let us note some of the simple mathematical properties of Eq. (7.12). If we carry out the point transformation

$$X \equiv R^{3/2}, \quad \Phi \equiv (\partial R / \partial X)^{1/2} \Psi = \left(\frac{2}{3}\right)^{1/2} R^{-1/4} \Psi, \quad (7.13)$$

Eq. (7.12) is converted to

$$-(3/64\pi^2) \partial^2 \Phi / \partial X^2 + 12\pi^2 X^{2/3} \Phi = Nm \Phi. \quad (7.14)$$

If the particles are in eigenstates of mass, so that  $m$  may be treated as a  $c$  number, and if the boundary condition

$$\Phi = 0 \text{ at } X = 0 \text{ or, equivalently, } \Psi = 0 \text{ at } R = 0 \quad (7.15)$$

analogous to (6.31) is imposed, then Eq. (7.14) becomes simply the Schrödinger equation of a particle of mass  $32\pi^2/3$  moving at energy  $Nm$  in the one-dimensional potential

$$\begin{aligned} V &= \infty, & X < 0, \\ V &= 12\pi^2 X^{2/3}, & X > 0. \end{aligned} \quad (7.16)$$

Now unless the mass eigenvalue  $m$  happens to be such that  $Nm$  is one of the allowed eigenvalues of Eq. (7.14), the function  $\Phi$  will not be normalizable but will behave in an exponential manner for large values of  $X$ . This is not necessarily bad if we insist on viewing  $R$ , and hence  $X$ , as a "time" coordinate, for it is usually impossible to require that a state function be normalizable with respect to *time*. Moreover, we may hesitate to allow the universe as a whole to determine the spectrum of masses which we can put into it, for in the classical theory the universe exerts no such control. However, several convincing arguments can be

adduced which suggest that  $\Phi$  must nonetheless be normalizable. The most important of these is that a closed Friedmann universe has, in the classical theory, a maximum radius of expansion. Hence if a correspondence principle is to exist, based on a transition to a classical limit,  $R$  must be effectively bounded from above. The existence of the classical turning point, as is well known, corresponds to the restriction to normalizable state functions.

For present purposes it suffices to determine the normalizable solutions of Eq. (7.12) in the WKB approximation. From the phase integral condition<sup>42</sup>

$$\begin{aligned} n + \frac{3}{4} &= -(2\pi)^{-1} \oint \Pi dR \\ &= 24\pi \int_0^{R_{\max}} [R(R_{\max} - R)]^{1/2} dR, \end{aligned} \quad (7.17)$$

$$R_{\max} \equiv Nm/12\pi^2, \quad (7.18)$$

we obtain the "energy" spectrum

$$Nm = [48\pi^2(n + \frac{3}{4})]^{1/2}, \quad n = 0, 1, 2, \dots \quad (7.19)$$

Computation of the normalized state function itself involves only elementary integrals. Inside the turning point it is found to have the form

$$\begin{aligned} \Psi &= (2/\pi R_{\max})^{1/2} [(R_{\max}/R) - 1]^{-1/4} \\ &\times \sin \{6\pi^2 [(2R - R_{\max})(R(R_{\max} - R))^{1/2} \\ &\quad + R_{\max}^2 \sin^{-1}((R/R_{\max})^{1/2})]\}, \end{aligned} \quad (7.20)$$

while outside it falls off to negligible values at distances of the order of  $R_{\max}^{-1/3}$  beyond  $R_{\max}$ .<sup>43</sup>

In realistic situations this function has an enormous number of nodes. For a Friedmann world approximating the actual universe one finds, very roughly,

$$n \sim 10^{120}, \quad (7.21)$$

and if all the degrees of freedom of the real world were taken into account the number would be vastly greater. However, despite the enormity of the quantum number, the function (7.20) does not provide a classical description of the universe, for it is a static function, composed of standing waves undergoing neither expansion nor contraction. The standing waves may, to be sure, be regarded as a superposition of waves "traveling" in opposite directions, those "traveling" in the direction of expansion (increasing  $R$ ) corresponding, by virtue of the "timelike" character of  $R$ , to the "positive frequency" components mentioned at the end of Sec. 5, and those "traveling" in the direction of contraction corresponding to "negative frequency"

<sup>42</sup> Here  $n + \frac{3}{4}$  is used in place of the usual  $n + \frac{1}{2}$  because of the "hard wall" character of the potential (7.16) for  $X < 0$ .

<sup>43</sup> Cf. Ref. 11, p. 462.

components.<sup>44</sup> However, in order to make this "travel" apparent we need some other coordinate besides  $R$ .

It is at this point that the internal particle dynamics enter the picture. The collective internal motion permits the particle ensemble to be used as a clock. Classically the temporal behavior of  $R$  may be determined by means of the correlation which exists between  $R$  and the  $q^i$ . This correlation is described by the solutions of the Hamilton-Jacobi equation

$$\mathfrak{H}(R, \partial W / \partial R) + \mathfrak{H}(q, \partial W / \partial q) = 0. \quad (7.22)$$

We may assume the particle Lagrangian  $l$  to be that of a multiply periodic system. The constants of integration of Eq. (7.22) are then conveniently taken to be the action variables  $\mathbf{J}_i$  of the collective Lagrangian  $L$ , and since the equation is obviously separable we have solutions of the form

$$W = -W(R, J) + W(q, \mathbf{J}) + \text{const.}, \quad (7.23)$$

where

$$J = -(2\pi)^{-1} \oint \Pi dR = J(\mathbf{J}), \quad \Pi = -\partial W / \partial R, \quad (7.24)$$

the integral being taken over a complete expansion-contraction cycle of the Friedmann universe. For these solutions Eq. (7.22) takes the separated form

$$\frac{1}{48\pi^2 R} \left( \frac{\partial W}{\partial R} \right)^2 + 12\pi^2 R = \mathfrak{H}(q, \partial W / \partial q) = E(\mathbf{J}), \quad (7.25)$$

where  $E(\mathbf{J})$  is a certain function of the  $\mathbf{J}_i$ .

The  $q^i$  are obtained as functions of  $R$  and the  $\mathbf{J}_i$  by integrating the simultaneous equations

$$dq^i / dR = V^i(q, \mathbf{J}) / V(R, J), \quad (7.26)$$

where

$$V^i = (\partial \mathfrak{H} / \partial P_i)_{P=\partial W / \partial q}, \quad (7.27)$$

$$V = (\partial \mathfrak{H} / \partial \Pi)_{\Pi=-\partial W / \partial R} = (24\pi^2 R)^{-1} \partial W / \partial R. \quad (7.28)$$

The integrals of Eqs. (7.26) are not hard to obtain. If Eq. (7.25) is differentiated with respect to  $\mathbf{J}_i$ , one finds

$$V \frac{\partial^2 W}{\partial R \partial J} \frac{\partial J}{\partial \mathbf{J}_i} = V^i \frac{\partial^2 W}{\partial q^i \partial \mathbf{J}_i} = \frac{\partial E}{\partial \mathbf{J}_i}, \quad (7.29)$$

which, together with (7.26), yields

$$\frac{\partial^2 W}{\partial J_i \partial q^j} dq^j = \frac{\partial J}{\partial \mathbf{J}_i} \frac{\partial^2 W}{\partial J \partial R} dR, \quad (7.30)$$

whence

$$-\frac{\partial W}{\partial J} \frac{\partial J}{\partial \mathbf{J}_i} + \frac{\partial W}{\partial \mathbf{J}_i} = \delta^i, \quad (7.31)$$

<sup>44</sup> Owing to the negative character of the kinetic-energy term of the Hamiltonian (7.10), the directions of "travel" of the exponential components of a standing wave are opposite to the conventional ones.

where the  $\delta^i$  are "phase constants." Equations (7.31) may be solved algebraically to express the  $q^i$ 's in terms of  $R$  and the constants of integration  $\mathbf{J}_i$ ,  $\delta^i$ . The  $\delta^i$  determine the relative phases of the simultaneous oscillatory motions, and the  $\mathbf{J}_i$  determine the amplitudes.

In the quantum theory an analogous correlation between  $R$  and the  $q^i$  can be established provided the state function has the form of a *superposition* of solutions of (7.12) corresponding to different eigenvalues of  $m$ . It is well known that a multiply periodic system cannot be used as a clock if it is in an eigenstate of energy. The uncertainty principle requires many different energy levels to be represented in its wave function. Here, however, we run into a very special difficulty which is peculiar to the quantum theory of space-time. The values which  $m$  can assume are already determined by the quantization condition (7.19) quite independently of the form of the particle Lagrangian  $l$ . Hence, unless the operators  $\mathfrak{H}$  and  $-\mathfrak{H}$  have at least one eigenvalue in common, the *Hamiltonian constraint* (7.1) will have no solutions at all. Equation (7.1) is unlike an ordinary time-independent Schrödinger equation in that it picks out only a single eigenvalue of the operator  $\mathfrak{H} + \mathfrak{H}$ . Moreover, the latter operator, being the sum of two operators having spectra bounded respectively from above and from below, has itself a spectrum which stretches from  $-\infty$  to  $\infty$ .

For purposes of the present discussion we must assume not only that  $\mathfrak{H} + \mathfrak{H}$  has a zero eigenvalue but that this eigenvalue is highly degenerate. We shall postpone until Sec. 10 a discussion of what the actual state of affairs may be in the real universe. For the present we concentrate on mathematical developments.

We shall confine ourselves to the WKB approximation and look for solutions of Eq. (7.1) of the form [cf. Eq. (6.11)]

$$\Psi = A \exp[i(-W + W)], \quad (7.32)$$

where  $A$  is a real amplitude satisfying (hopefully) the inequalities [cf. Eq. (6.12)]

$$\left| \frac{\partial A}{\partial R} \right| \ll \left| A \frac{\partial W}{\partial R} \right|, \quad \left| \frac{\partial A}{\partial q^i} \right| \ll \left| A \frac{\partial W}{\partial q^i} \right|. \quad (7.33)$$

A differential equation for  $A$  may be obtained by substituting (7.32) into (7.1). One finds

$$[\mathfrak{H}(R, -i\partial/\partial R - \partial W / \partial R) + \mathfrak{H}(q, -i\partial/\partial q + \partial W / \partial q)]A = 0. \quad (7.34)$$

When the inequalities (7.33) are satisfied the "big" terms of (7.34) already add up to zero by virtue of the Hamiltonian-Jacobi equation (7.22). In order to obtain conditions on  $A$  we must include the smaller, "higher-order" terms, and for this purpose it is convenient to introduce a smooth real test function  $\varphi(R, q)$  which vanishes outside a finite closed region in the  $R-q$

manifold.<sup>45</sup> We then subtract the equation

$$\int_0^\infty dR \int dq \varphi A [\Im C(R, -i\partial/\partial R - \partial W/\partial R) + \Re C(q, -i\partial/\partial q + \partial W/\partial q)] A = 0 \quad (7.35)$$

from its complex conjugate and use the Hermiticity of  $\Im C$  and  $\Re C$  to obtain

$$\int_0^\infty dR \int dq A [\varphi, \Im C(R, -i\partial/\partial R - \partial W/\partial R) + \Re C(q, -i\partial/\partial q + \partial W/\partial q)] A = 0. \quad (7.36)$$

When the inequalities (7.33) are satisfied, this becomes approximately

$$i \int_0^\infty dR \int dq [(\partial \varphi / \partial R) V + (\partial \varphi / \partial q^i) V^i] A^2 = 0, \quad (7.37)$$

which, by virtue of the arbitrariness of  $\varphi$ , implies (after an integration by parts)

$$\partial(A^2 V) / \partial R + \partial(A^2 V^i) / \partial q^i = 0. \quad (7.38)$$

This equation, which is the analog of (6.14), assures conservation of the "flux of probability" in the  $R$ - $q$  manifold.

The general solution of Eq. (7.38) can be obtained by making use of the relations

$$V \partial \delta^i / \partial R + V^j \partial \delta^i / \partial q^j = 0, \quad (7.39)$$

$$\partial(V \partial^2 W / \partial R \partial J) / \partial R = 0, \quad (7.40)$$

$$\partial(V^i D) / \partial q^i = 0, \quad (7.41)$$

where

$$D \equiv \det(D_{ij}), \quad D_{ij} = \partial^2 W / \partial q^i \partial J_j, \quad (7.42)$$

and where  $\delta^i$ , in Eq. (7.39) is regarded not as a constant of integration but as a function of  $R$  and the  $q^i$ , defined by (7.31). It is easy to see that Eq. (7.39) follows from (7.29) and (7.31). Equation (7.40) is obtained by differentiating (7.29) with respect to  $R$ , while (7.41) is obtained by differentiating (7.29) with respect to  $q^k$  and multiplying by the matrix  $D^{-1}_{ik}$  inverse to (7.42). From these relations it follows immediately that

$$A^2 = (\partial^2 W / \partial R \partial J) DF(\delta), \quad (7.43)$$

where  $F$  is an arbitrary function of the  $\delta^i$ .

Actually the form of  $F$  is not arbitrary, since there are other differential equations which the state function (7.32) must satisfy in addition to (7.12), namely, the eigenvalue equations

$$J(R, -i\partial/\partial R) \Psi = J\Psi, \quad (7.44)$$

$$\mathbf{J}_i(q, -i\partial/\partial q) \Psi = \mathbf{J}_i\Psi. \quad (7.45)$$

<sup>45</sup> It is always to be understood that the  $R$ - $q$  manifold is restricted to positive values of  $R$  (excluding zero).

The operators  $J(R, -i\partial/\partial R)$  and  $\mathbf{J}_i(q, -i\partial/\partial q)$  are obtained by solving the equations

$$\Pi = -\partial W / \partial R, \quad P_i = \partial W / \partial q^i \quad (7.46)$$

for the  $J$ 's in terms of  $\Pi$ ,  $R$ , and the  $P$ 's and  $q$ 's, making the replacements  $\Pi \rightarrow -i\partial/\partial R$ ,  $P_i \rightarrow -i\partial/\partial q^i$ , and carrying out appropriate Hermiticity symmetrizations. Now

$$J(R, -i\partial/\partial R) \Psi = \{\exp[i(-W + \mathbf{W})]\} \times J(R, -i\partial/\partial R - \partial W / \partial R) A, \quad (7.47)$$

$$\mathbf{J}_i(q, -i\partial/\partial q) \Psi = \{\exp[i(-W + \mathbf{W})]\} \times \mathbf{J}_i(q, -i\partial/\partial q + \partial W / \partial q) A. \quad (7.48)$$

Because of the identities

$$J(R, -\partial W / \partial R) \equiv J, \quad \mathbf{J}_i(q, \partial W / \partial q) \equiv \mathbf{J}_i \quad (7.49)$$

the "big" terms of (7.47) and (7.48) already yield Eqs. (7.44) and (7.45). Hence the "higher-order" terms must vanish. With the aid of the inequalities (7.33) and a test function  $\varphi$ , as before, one easily finds that this implies

$$\partial(A^2 \partial J / \partial R) / \partial R = 0, \quad (7.50)$$

$$\partial(A^2 \partial \mathbf{J}_i / \partial P_i) / \partial q^i = 0. \quad (7.51)$$

But  $\partial J / \partial \Pi = -(\partial^2 W / \partial R \partial J)^{-1}$  and  $\partial \mathbf{J}_i / \partial P_i = D^{-1}_{ii}$ . Hence, substituting (7.43) into (7.47) and (7.48), and making use of the identity

$$\partial(DD^{-1}_{ii}) / \partial q^j = 0. \quad (7.52)$$

which can be shown to hold by virtue of the symmetry of  $\partial D_k^l / \partial q^j$  in  $j$  and  $k$ , one finds that  $F$  must be a constant, independent of the  $\delta^i$ .

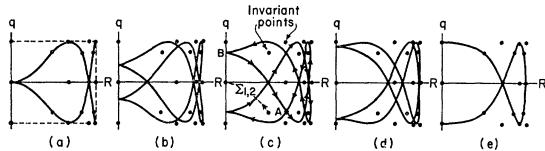
In order to obtain normalizable solutions of (7.12) the  $J$ 's must be quantized. In addition, the branch-point behavior of the functions  $W$  and  $\mathbf{W}$  at the classical turning points must be taken into account, and superpositions of the form (7.32) corresponding to the different branches must be employed. These superpositions are the standard WKB solutions.

Suppose we freeze out all the collective particle degrees of freedom save one, and suppose this degree of freedom corresponds to a motion of libration in a smooth potential. Then we may distinguish two branches of the function  $\mathbf{W}$ , a branch  $\mathbf{W}^+$  which increases with increasing  $q$  and a branch  $\mathbf{W}^-$  which decreases with the increasing  $q$ . Similarly we may distinguish two branches,  $W^+$  and  $W^-$ , of the function  $W$ . These branches are determined only up to arbitrary constants. The constants may be adjusted so that the WKB solutions take the form

$$\Psi_n = A_n [\exp(-iW_n^+) + \exp(-iW_n^-)] \times [\exp(iW_n^+) + \exp(iW_n^-)], \quad (7.53)$$

where

$$W_n^\pm \equiv W^\pm(R, n + \frac{3}{4}), \quad W_n^\pm \equiv W^\pm(q, n + \frac{1}{2}), \quad (7.54)$$

FIG. 1. Packet traces in the  $R$ - $q$  plane. (Case  $\Delta n=3$ ,  $\Delta \mathbf{n}=2$ .)

only those quantum numbers  $n$ ,  $\mathbf{n}$  being permitted for which

$$J = n + \frac{3}{4} \text{ when } \mathbf{J} = \mathbf{n} + \frac{1}{2}. \quad (7.55)$$

If the operator  $\mathcal{H} + \mathcal{C}$  is to have a zero eigenvalue which is highly degenerate, it is clear from Eqs. (7.25) and (7.55) that the coherent internal dynamical behavior of the particles filling the universe must be precisely matched to that of the universe itself in such a way that the derivative  $dJ/d\mathbf{J}$  is a *constant rational number* over a wide range of values of  $\mathbf{J}$ . We shall write

$$dJ/d\mathbf{J} = \Delta n/\Delta \mathbf{n}, \quad (7.56)$$

where  $\Delta n$  and  $\Delta \mathbf{n}$  are relatively prime integers.  $\Delta n$  and  $\Delta \mathbf{n}$  are the spacings between adjacent permitted values of the quantum numbers  $n$  and  $\mathbf{n}$ , respectively. An immediate consequence of Eq. (7.56) is that the angular frequencies (with respect to proper time) of the  $R$  and  $q$  motions are always (within the allowed range of  $\mathbf{J}$  values) commensurable. These angular frequencies are given by

$$\omega \equiv dE/dJ = \omega d\mathbf{J}/dJ, \quad (7.57)$$

$$\omega \equiv dE/d\mathbf{J}, \quad (7.58)$$

and, in the allowed range, satisfy

$$\omega \Delta n = \omega \Delta \mathbf{n}. \quad (7.59)$$

Use of the angular frequencies permits Eqs. (7.29) to be re-expressed in the form

$$V \frac{\partial^2 W}{\partial R \partial J} = \omega, \quad V \frac{\partial^2 W}{\partial q \partial \mathbf{J}} = \omega, \quad (7.60)$$

whence (7.43) becomes

$$A^2 = F \omega \omega / VV. \quad (7.61)$$

For later convenience we shall choose

$$F = \Delta n / 2\pi\omega = \Delta \mathbf{n} / 2\pi\omega. \quad (7.62)$$

The normalization constant  $A_n$  in expression (7.53) is then given by

$$A_n = (\omega_n \Delta n / 2\pi |V_n V_{\mathbf{n}}|)^{1/2} \approx [(E_{n+\Delta n} - E_n) / 2\pi |V_n V_{\mathbf{n}}|]^{1/2}, \quad (7.63)$$

the subscripts indicating that the quantized values of the quantities to which they are affixed are to be employed. No signs have been placed on the  $V$ 's to correspond with the different branches of the  $W$ 's, because

motion in a potential is time-reversal invariant, and hence  $|V^+| = |V^-|$ ,  $|V^+| = |V^-|$ .

### 8. A WAVE PACKET FOR THE UNIVERSE. THE CONCEPT OF TIME

We are now in a position to construct a state function exhibiting classical behavior. We do this by superposing many WKB solutions:

$$\Psi_a = \sum_{n'} a'_n \Psi_n, \quad (8.1)$$

the prime indicating that the summation is to be carried out only over those quantum numbers which satisfy condition (7.55). If the  $a$ 's are carefully chosen,  $\Psi_a$  will have the form of a "wave packet" which traces out a classical trajectory, namely, a generalized lissajous figure in the  $R$ - $q$  plane. It is easy to see that Eq. (7.31) is just the condition for constructive interference, provided the symbols in the equation are understood to denote the "peak" values of the quantities to which they refer.

In view of the commensurability condition (7.59) the lissajous figure is necessarily closed and of finite length.<sup>46</sup> A typical set of packet traces is shown in Fig. 1 for the case  $\Delta n=3$ ,  $\Delta \mathbf{n}=2$ . The action variable  $\mathbf{J}$  is the same for each trace, but the phase constant  $\delta$  varies from one to the other. The following facts may be inferred from the figures: Each trace lies within a rectangle having sides equal to the full amplitudes of the oscillations. Except in the degenerate cases depicted in Figs. 1(a) and 1(e), each trace divides the rectangle into  $2\Delta n \Delta \mathbf{n} + \Delta n + \Delta \mathbf{n} + 1$  disjoint regions. Although the size and shape of corresponding regions vary from figure to figure, each region contains an *invariant point* which is independent of  $\delta$ . These points are shown in the figures.

The degenerate curves are those which have collapsed onto the invariant points. They are divided by the invariant points into a total of  $4\Delta n \Delta \mathbf{n}$  segments, which will be called *invariant segments*. The invariant segments may be labeled in a systematic fashion, starting, say, from the "southwest" corner of the enclosing rectangle, by pairs of integers  $(r, r)$  satisfying  $1 \leq r \leq 2\Delta n$ ,  $1 \leq r \leq 2\Delta \mathbf{n}$ . When an invariant segment is used as a contour of integration (see Sec. 9) it will be denoted by the symbol  $\Sigma_{r,r}$ . [See Fig. 1(c).] Each invariant segment corresponds to a lapse of proper time of amount  $T/4\Delta n = \bar{T}/4\Delta \mathbf{n}$  where  $T$  and  $\bar{T}$  are the oscillation periods.

<sup>46</sup> Degenerate forms of closure [see, for example, Figs 1(a) and 1(e)], in which the packet "moves" back and forth along the same curve, are to be understood as included in this statement. Also, if successive groups of  $a$ 's are chosen to vanish in such a way that the effective spacing between adjacent quantum numbers becomes a multiple of  $\Delta n$  (or  $\Delta \mathbf{n}$ ), then the packet trace will consist of  $m$  separately closed Lissajous figures superimposed upon one another. Such a trace must be understood as representing a *single* packet which consists of  $m$  disconnected parts. Although the following discussion can be extended to include such situations, they will, for simplicity, be excluded from consideration.

A wave packet will be called *good* if its state function has negligible values throughout most of each region containing an invariant point. Except at intersection points or turning points only one of the branches of each of the functions  $W$  and  $\mathbf{W}$  is involved in the constructive interference of the WKB functions at any one position along a good packet trace. Thus, for "motion" in a "northeasterly" direction the relevant branches are  $W^+$ ,  $\mathbf{W}^+$ , provided we use the standard convention that time increases in the direction of increasing  $W$  and  $\mathbf{W}$ . Continuing around the compass we have  $W^+$ ,  $\mathbf{W}^-$  for SE;  $W^-$ ,  $\mathbf{W}^-$  for SW; and  $W^-$ ,  $\mathbf{W}^+$  for NW. The trace is thus divided into *branch segments* having definite quadrant orientations.

A given branch segment may be intersected by other branch segments which further subdivide it. The resulting pieces will be called *simple segments*. Each simple segment is intersected by precisely one invariant segment, and the two may therefore be labeled by the same integers. The branches involved in a given simple segment are  $W^+$ ,  $\mathbf{W}^+$ , or  $W^-$ ,  $\mathbf{W}^-$  if  $r+r$  is odd and  $W^+$ ,  $\mathbf{W}^-$ , or  $W^-$ ,  $\mathbf{W}^+$  if  $r+r$  is even, the choice depending on the direction of "motion." The direction in which the packet moves as time increases may be indicated by affixing arrows to the packet trace, as in Fig. 1(c). If the  $W$ 's are adjusted so that  $\delta=0$  corresponds to a degenerate trace, then the arrows are reversed by changing the sign of  $\delta$ .

Proper time itself is defined by

$$\tau = \omega^{-1}\Theta, \quad \tau = \omega^{-1}\Theta, \quad (8.2)$$

where the  $\omega$ 's are defined by Eqs. (7.57) and (7.58), and

$$\Theta = \partial W / \partial J, \quad \Theta = \partial \mathbf{W} / \partial \mathbf{J}. \quad (8.3)$$

Classically the angle variables  $\Theta$  and  $\Theta$  are canonically conjugate to  $-J$  and  $J$ , respectively, and hence  $\tau$  and  $\tau$  are canonically conjugate to  $\mathcal{H}$  and  $\mathcal{H}$ , respectively. In the quantum theory this leads one to write the commutation relations

$$[\tau, \mathcal{H}] = i, \quad [\tau, \mathcal{H}] = i. \quad (8.4)$$

It is important to remember, however, that the quantum  $\tau$ 's are *not Hermitian*. This follows not only from their periodic character, which arises from their dependence on the  $\Theta$ 's [Eqs. (8.2)], but also from the fact that their canonical conjugates,  $\mathcal{H}$  and  $\mathcal{H}$ , have discrete, "one-sided" spectra, bounded, respectively, from above and below. The usual eigenvector properties which hold for Hermitian operators therefore do not hold for the  $\tau$ 's, and we must distinguish between right and left eigenvectors.

Let us introduce the left eigenvectors  $\langle \tau', \tau' |$  which, in virtue of (8.4), may be chosen to satisfy

$$-i\frac{\partial}{\partial \tau'} \langle \tau', \tau' | = \langle \tau', \tau' | \mathcal{H}, \quad -i\frac{\partial}{\partial \tau'} \langle \tau', \tau' | = \langle \tau', \tau' | \mathcal{H}. \quad (8.5)$$

We may also introduce the corresponding conjugate vectors, denoted by  $| \tau', \tau' \rangle$ , which are right eigenvectors of the conjugate operators  $\tau^\dagger$  and  $\tau^\dagger$ . Now let  $\Phi$  denote the projection operator into the physical subspace of allowed state vectors. Using Eqs. (8.5) and the Hamiltonian constraint (7.1), which may be rewritten in the form

$$(\mathcal{H} + \mathcal{H})\Phi = \Phi(\mathcal{H} + \mathcal{H}) = 0, \quad \Phi^2 = \Phi, \quad (8.6)$$

it is easy to see that  $\langle \tau, \tau | \Phi$  depends only on the difference  $\tau - \tau$ . This simple dependence may be recognized as a quantum consequence of the classical correlation

$$\tau - \tau = -\omega^{-1}\delta, \quad (8.7)$$

which follows from (7.31) and (8.2).

The projection operator  $\Phi$  is conveniently defined in terms of the eigenvectors  $| n + \frac{3}{4}, n + \frac{1}{2} \rangle$  of the  $J$ 's. Writing  $| n + \frac{3}{4}, n + \frac{1}{2} \rangle \equiv | n \rangle$  whenever the quantum numbers are restricted as in (7.55), we have

$$\Phi = \sum_n' | n \rangle \langle n |, \quad (8.8)$$

provided the normalization  $\langle n | n' \rangle = \delta_{nn'}$  is assumed. In virtue of Eqs. (8.5) we may also write

$$\langle \tau, \tau | n \rangle = (\omega_n \Delta n / 2\pi)^{1/2} \exp[-iE_n(\tau - \tau)]. \quad (8.9)$$

The normalization here is chosen so as to maximize the orthogonality properties of the vectors  $\langle \tau, \tau |$  relative to the physical subspace. Noting that  $\omega_n \Delta n$  gives the approximate spacing between adjacent permitted "energy" eigenvalues  $E_n$ , we have

$$\begin{aligned} \langle \tau, \tau | \Phi | \tau', \tau' \rangle &\approx (2\pi)^{-1} \sum_n' (\exp\{-iE_n[(\tau - \tau) - (\tau' - \tau')]\}) \Delta E_n \\ &= \delta((\tau - \tau) - (\tau' - \tau')). \end{aligned} \quad (8.10)$$

If the "energy" spectrum were continuous and ranged from  $-\infty$  to  $\infty$  the function  $\delta$  would be the Dirac  $\delta$ . In reality it is a function which although divergent at the origin does not completely vanish elsewhere. Thus the eigenvectors  $\langle \tau, \tau |$  are only approximately orthonormal, a fact which stems from the lack of strict Hermiticity of the operators  $\tau$  and  $\tau$ .

Instead of working with the vectors  $\langle \tau, \tau |$  it is more interesting to work with  $\langle \tau, q |$ ,  $\langle R, \tau |$ , and  $\langle R, q |$ , which are defined in an obvious fashion. The normalization of  $\langle R, q |$  may be fixed by setting

$$\langle R, q | n \rangle = \Psi_n, \quad (8.11)$$

where  $\Psi_n$  is the function having the WKB approximation (7.53), with  $A_n$  given by (7.63). In a similar manner the normalization of  $\langle \tau, q |$  and  $\langle R, \tau |$  may be fixed by giving the WKB approximations of their inner products with  $| n \rangle$ . We shall choose

$$\begin{aligned} \langle \tau, q | n \rangle &\equiv \Phi_n = (\omega_n \Delta n / 2\pi | V_n |)^{1/2} \\ &\times e^{-iE_n \tau} [\exp(iW_n^+) + \exp(iW_n^-)], \end{aligned} \quad (8.12)$$

$$\begin{aligned} \langle R, \tau | n \rangle &\equiv \Phi_n = (\omega_n \Delta n / 2\pi | V_n |)^{1/2} [\exp(-iW_n^+) \\ &+ \exp(-iW_n^-)] \exp(iE_n \tau). \end{aligned} \quad (8.13)$$

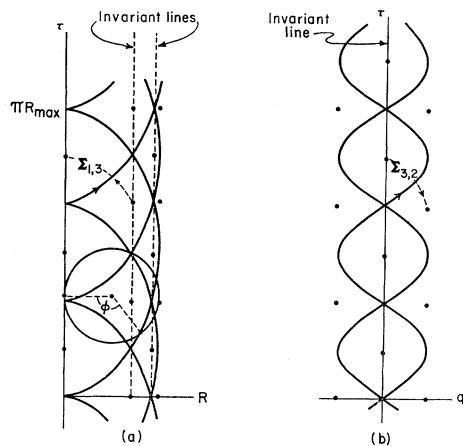


FIG. 2. Packet traces in the (a)  $R-\tau$  and (b)  $\tau-q$  planes. (Case  $\Delta n=3$ ,  $\Delta n=2$ .)

Denoting by  $\Phi$  and  $\Psi$  arbitrary superpositions of the functions  $\Phi_n$  and  $\Psi_n$ , respectively, we may write the Schrödinger equations

$$i\partial\Phi/\partial\tau = \mathcal{H}(q, -i\partial/\partial q)\Phi, \quad (8.14)$$

$$i\partial\Psi/\partial\tau = \mathcal{H}(R, -i\partial/\partial R)\Psi. \quad (8.15)$$

When the function  $\Psi_a$  of (8.1) has the form of a wave packet so also have the corresponding functions  $\Phi_a$  and  $\Psi_a$ . The form of the packet trajectory in the case of the function  $\Phi_a$  may be determined by noting that the condition for constructive interference, which establishes the correlation between  $R$  and  $\tau$ , is

$$\partial W/\partial J = \omega\tau + \delta, \quad (8.16)$$

where  $\delta$  is a phase constant and the other symbols denote the peak values of the quantities to which they refer.<sup>47</sup> Differentiation of Eq. (8.16) with respect to  $\tau$  and use of Eqs. (7.28) and (7.60) yields

$$\begin{aligned} dR/d\tau &= V = -(24\pi^2 R)^{-1}\Pi \\ &= R^{-1}[R(R_{\max}-R)]^{1/2}. \end{aligned} \quad (8.17)$$

The integration of this equation is most easily carried out with the aid of an angle  $\phi$  defined by

$$d\tau/d\phi = R. \quad (8.18)$$

This gives

$$d\phi = [R(R_{\max}-R)]^{-1/2}dR, \quad (8.19)$$

which yields the familiar cyclodial trajectory of the dust-filled Friedmann universe:

$$R = \frac{1}{2}R_{\max}(1-\cos\phi), \quad (8.20)$$

$$\tau = \frac{1}{2}R_{\max}(\phi - \sin\phi), \quad (8.21)$$

the constants of integration being chosen so that  $R=0$ ,  $\phi=0$  at  $\tau=0$ . We note that by virtue of the boundary

<sup>47</sup> Equation (8.16) also follows from (8.2), (8.3), and (8.7).

condition (7.15) the packet rebounds repeatedly from the collapsed state until it ultimately loses its identity owing to spreading. Throughout the period of each rebound the width of the packet remains at all times finite, never suffering infinite compression. Transition through collapse thus becomes, in the quantum theory, a continuous process—something which cannot be achieved within the classical framework.

Figure 2 shows the curves traced out in the  $R-\tau$  and  $\tau-q$  planes by the packet of Fig. 1, and reveals a slight complication which was overlooked in the above simple analysis. The curves in the two planes appear to depict  $\Delta n$  and  $\Delta n$  distinct packets, respectively, rather than only a single packet. The extra “ghost packets” arise because the complete spectra of  $\mathcal{C}$  and  $\mathcal{C}$  are not made use of in the superposition (8.1). Only every  $\Delta n$ th level of  $\mathcal{C}$  and every  $\Delta n$ th level of  $\mathcal{C}$  occur. This means that  $\tau$  and  $\tau$  are determined modulo  $T/\Delta n = T/\Delta n$  rather than modulo  $T$  and  $T$ , respectively, and a given packet must consequently appear “simultaneously” in several places in order to allow for the resulting proper time ambiguity.<sup>48</sup>

The multiple traces intersect themselves along  $\Delta n-1$  and  $\Delta n-1$  phase-invariant lines, respectively. (See the indicated lines in the figures.) These lines divide the branch segments (which are defined just as for the  $R-q$  lissajous figures) into simple segments. Each simple segment is straddled by a unique pair of points at which maximum destructive interference occurs. The pairs of points may be connected by arcs intersecting the associated simple segments. These arcs will be denoted by  $\Sigma_{r,t}$  and  $\Sigma_{t,r}$  in the  $R-\tau$  and  $\tau-q$  planes, respectively. The pairs of suffixes  $r$ ,  $t$  and  $t$ ,  $r$  have the ranges  $r=1, 2, \dots, \Delta n; t=1, 2, \dots, \Delta n; t=-2, -1, 0, 1, 2, \dots$ , and may be used in an obvious manner to identify either the simple segments or their associated intersecting arcs. Examples of arcs and point pairs are shown in the figure.

With the introduction of the three wave functions  $\Phi$ ,  $\Psi$ , and  $\Psi$  we now have at our disposal three distinct mathematical windows from which to view the Friedmann world. From one window the material content of the universe is seen as a clock for determining the dynamical behavior of the world geometry. From another it is the geometry which appears as a clock for determining the dynamical behavior of the material content. From the third the geometry and the material content appear on equal footing, each one correlated in a certain manner with the other.

It is the third window which is to be preferred as most accurately revealing the physics of the quantized Friedmann model. The variables  $\tau$  and  $\tau$ , because of their lack of Hermiticity, are not rigorously observable and hence cannot yield a measure of proper time which is valid under all circumstances. It is only with good

<sup>48</sup> This has also the consequence that  $\langle R, \tau | \Phi | R', \tau' \rangle$  and  $\langle \tau, q | \Phi | \tau', q' \rangle$  do not have the form of simple  $\delta$  functions, although they diverge at  $R=R'$  and  $q=q'$ .

wave packets that these variables are useful. But even with a good packet the description in terms of  $\tau$  and  $\pi$  is not perfect, as is revealed in a striking way by the fact that the wave packets  $\Phi_a$  and  $\Phi_a$  inevitably spread in "time," whereas the packet  $\Psi_a$  does not. It is for this reason that we may say that "time" is only a phenomenological concept, useful under certain circumstances.

It is worth remarking that it is not necessary to drag in the whole universe to argue for the phenomenological character of time. If the principle of general covariance is truly valid then the quantum mechanics of every-day usage, with its dependence on Schrödinger equations of the form (8.14) or (8.15), is only a phenomenological theory. For the only "time" which a covariant theory can admit is an intrinsic time defined by the contents of the universe itself. Any intrinsically defined time is necessarily non-Hermitian, which is equivalent to saying that there exists no clock, whether geometrical or material, which can yield a measure of time which is operationally valid under *all* circumstances, and hence there exists no operational method for determining the Schrödinger state function with arbitrarily high precision. This statement also follows directly from the uncertainty principle. Because every clock has a "one-sided" energy spectrum, its ultimate accuracy must necessarily be inversely proportional to its rest mass. When the whole universe is cast in the role of a clock, the concept of time can of course be made fantastically accurate (at least in principle) because of the enormity of the masses and quantum numbers involved. But as long as the universe is finite, a theoretical limit to the accuracy nevertheless remains.

## 9. THE INNER PRODUCT

We shall now use the results of the two preceding sections to show how the definition (5.19) for inner products can be rescued from the negative-probability disaster, at least in the case of the quantized Friedmann model. First we must derive the form which (5.19) takes in this model. Consider the following integral:

$$\int \varphi \{ [\mathfrak{J}\mathcal{C}(q, -i\partial/\partial q)\Psi_b]^* \Psi_a - \Psi_b^* \mathfrak{J}\mathcal{C}(q, -i\partial/\partial q)\Psi_a \} dq,$$

where  $\Psi_a$  and  $\Psi_b$  are arbitrary complex functions of  $R$  and the  $q^i$ , and  $\varphi$  is a real test function. Because of the Hermiticity of  $\mathfrak{J}\mathcal{C}$  this integral may be rewritten in the form

$$\begin{aligned} & \int \Psi_b^* [\mathfrak{J}\mathcal{C}(q, -i\partial/\partial q), \varphi] \Psi_a dq \\ &= -i \int \Psi_b^* \mathbf{V}^i(q, -i\partial/\partial q) \bullet (\partial \varphi / \partial q^i) \Psi_a dq, \end{aligned} \quad (9.1)$$

where  $\mathbf{V}^i$  is defined by (7.27), but with the replacement  $P_i \rightarrow -i\partial/\partial q^i$  instead of  $P_i = \partial W / \partial q^i$ , and where the

dot in the right-hand integrand indicates that the factor  $\partial \varphi / \partial q^i$  is to be inserted between noncommuting factors in the terms of  $\mathbf{V}^i$  in such a way as to yield the commutator on the left. If now the differential operators occurring in  $\mathbf{V}^i$  are peeled to the left and right, via integrations by parts, in such a manner that they no longer act on  $\partial \varphi / \partial q^i$ , then the integral takes the form

$$\begin{aligned} & -i \int (\partial \varphi / \partial q^i) (\Psi_b^* \overset{\leftrightarrow}{\mathbf{V}}^i \Psi_a) dq \\ &= i \int \varphi \partial (\Psi_b^* \overset{\leftrightarrow}{\mathbf{V}}^i \Psi_a) / \partial q^i dq, \end{aligned} \quad (9.2)$$

where  $\overset{\leftrightarrow}{\mathbf{V}}^i$  denotes the result of the peeling process. Because of the arbitrariness of  $\varphi$  it follows that

$$\begin{aligned} & [\mathfrak{J}\mathcal{C}(q, -i\partial/\partial q)\Psi_b]^* \Psi_a - \Psi_b^* \mathfrak{J}\mathcal{C}(q, -i\partial/\partial q)\Psi_a \\ &= \partial (\Psi_b^* \overset{\leftrightarrow}{\mathbf{V}}^i \Psi_a) / \partial q^i. \end{aligned} \quad (9.3)$$

In a similar manner we find

$$\begin{aligned} & [\mathfrak{J}\mathcal{C}(R, -i\partial/\partial R)\Psi_b]^* \Psi_a - \Psi_b^* \mathfrak{J}\mathcal{C}(R, -i\partial/\partial R)\Psi_a \\ &= \partial (\Psi_b^* \overset{\leftrightarrow}{\mathbf{V}}^i \Psi_a) / \partial R, \end{aligned} \quad (9.4)$$

where in this case we can give an explicit form for  $\overset{\leftrightarrow}{\mathbf{V}}$ :

$$\overset{\leftrightarrow}{\mathbf{V}} = \frac{i}{48\pi^2} \left( R^{-3/4} \frac{\partial}{\partial R} R^{-1/4} - R^{-1/4} \frac{\partial}{\partial R} R^{-3/4} \right). \quad (9.5)$$

The analog of (5.19) is now obvious, namely,

$$(\Psi_b, \Psi_a) = \int_{\Sigma} (\Psi_b^* \overset{\leftrightarrow}{\mathbf{V}} \Psi_a d\Omega + \Psi_b^* \overset{\leftrightarrow}{\mathbf{V}}^i \Psi_a dR d\Sigma_i), \quad (9.6)$$

where  $\Sigma$  is an appropriate surface in the  $R$ - $q$  manifold and  $d\Sigma_i$  is the directed surface element of its projection into  $q$  space. From Eqs. (9.3) and (9.4) it follows that

$$\partial (\Psi_b^* \overset{\leftrightarrow}{\mathbf{V}} \Psi_a) / \partial R + \partial (\Psi_b^* \overset{\leftrightarrow}{\mathbf{V}}^i \Psi_a) / \partial q^i = 0, \quad (9.7)$$

whenever  $\Psi_a$  and  $\Psi_b$  are physical state functions satisfying the Hamiltonian constraint (7.1). Therefore the integral (9.6) is independent of  $\Sigma$  provided the boundary of  $\Sigma$  remains in a region where  $\Psi_a$  and  $\Psi_b$  vanish.

When the coherent dust filling the Friedmann universe is restricted to only one degree of freedom the inner product (9.6) reduces to

$$(\Psi_b, \Psi_a) = \int_{\Sigma} (\Psi_b^* \overset{\leftrightarrow}{\mathbf{V}} \Psi_a d\Omega - \Psi_b^* \overset{\leftrightarrow}{\mathbf{V}}^i \Psi_a dR), \quad (9.8)$$

where  $\Sigma$  is an appropriate contour in the  $R$ - $q$  plane. The key word here is "appropriate." In analogy with our previous treatment of the manifold  $\mathfrak{M}$  of 3-geometries in the general theory, we may view the  $R$ - $q$  plane as endowed with a natural metric determined by the structure of the functions  $\mathfrak{J}\mathcal{C}$  and  $\mathfrak{J}\mathcal{C}$ . With respect to this metric the coordinates  $R$  and  $q$  are "timelike" and "spacelike", respectively. If the Hamiltonian constraint

(7.1) were an ordinary wave equation we would naturally adopt for the contour  $\Sigma$  a "spacelike" line such as  $R=\text{constant}$ . However, just as in the general theory, so also here, "wave" propagation is not restricted to timelike directions. Indeed, from the lissajous traces of Fig. 1, it is evident that the Friedmann universe not only executes "timelike" and "spacelike" motions with impartiality, but even turns around and "moves" backward with respect to the "time" coordinate. The distinction between "timelike" and "spacelike" clearly does not have the same pervasive significance here as it does in ordinary wave theories.

If we were actually to choose, for  $\Sigma$ , a line  $R=\text{constant}$ , we would obtain the useless result  $(\Psi_b, \Psi_a) = 0$  for all  $\Psi_a$  and  $\Psi_b$ . This is because all physically admissible state functions have non-negligible values only in a finite domain of the  $R$ - $q$  plane. Hence any line  $R=\text{constant}$  can be deformed into a line along which  $\Psi_a$  and  $\Psi_b$  effectively vanish, without affecting the value of the integral (9.8) at all. The same is true if  $\Sigma$  is a "timelike" curve which starts at  $R=0$  and goes out to infinity in the  $R$ - $q$  plane. Since any normalizable superposition of the functions (7.20) vanishes at  $R=0$  at least as fast as  $R^{7/4}$ , such a curve can also be deformed into one along which  $\Psi_a$  and  $\Psi_b$  vanish, without affecting (9.8).

How then shall we choose  $\Sigma$ ? The answer is to be found in the conservation laws (6.13), (7.38), and (9.7). From our analysis of the Lissajous traces of Fig. 1 it is evident that probability flows in a closed finite circuit in the  $R$ - $q$  plane.  $\Sigma$  must therefore be a *finite* curve, chosen so as to intersect a unidirectional unit flux of probability of each of the two functions  $\Psi_a$  and  $\Psi_b$ .

This means that Eq. (9.8) can be used to define inner products only when  $\Psi_a$  and  $\Psi_b$  both have the form of good packets. If they do not have this form or if they fall into the degenerate category depicted in Figs. 1(a) and 1(e), then some other representation must be employed. An analogous condition must hold in the general theory if Eq. (5.19) is to be valid. Whenever the condition is violated the usefulness of Wheeler's "metric representation" diminishes.

It is not difficult to show that Eq. (9.8) indeed yields an acceptable value for the inner product under the required conditions. The case in which the two packets do not overlap (except at intersections) may be disposed of at once;  $(\Psi_b, \Psi_a)$  then clearly vanishes. The only case which need concern us is that in which the two packets overlap, at least partially, throughout the entire length of their trajectories. The  $\mathbf{J}$  values at their "peaks" then differ negligibly compared to the spread of values contained in their superpositions. As our initial contour we shall choose an invariant segment  $\Sigma_{r,r}$  corresponding to the average of the peak  $\mathbf{J}$  values. Since the packets are "good" we know that  $\Psi_a$  and  $\Psi_b$  vanish at its endpoints. Suppose  $\Sigma_{r,r}$  intersects a NE-SW branch of the lissajous figure formed by the packet traces [e.g., the segment  $\Sigma_{1,2}$  shown in Fig. 1(c)], and suppose the orientation of both packets is the same, say NE, at the point of intersection with  $\Sigma_{r,r}$ . Then it is only the  $W^+$ ,  $W^+$  branch of each packet which interferes constructively along  $\Sigma_{r,r}$ . Approximating  $\Psi_a$  and  $\Psi_b$  by their WKB forms, keeping only those parts which refer to the branch in question, and taking note of the inequalities (7.33), we have

$$\begin{aligned} (\Psi_b, \Psi_a) &= \int_{\Sigma_{r,r}} (\Psi_b * \vec{\nabla} \Psi_a dq - \Psi_b * \vec{\nabla} \Psi_a) dR \\ &= \sum_{n,n'} b_{n'} * a_n \int_{\Sigma_{r,r}} (\omega_{n'} \omega_n \Delta n' \Delta n / 4\pi^2 |V_{n'} V_n V_{n'} V_n|)^{1/2} \{ [\exp(iW_{n'}^+) \vec{\nabla} \exp(-iW_n^+)] \\ &\quad \times \exp[-i(W_{n'}^+ - W_n^+)] dq - \exp[i(W_{n'}^+ - W_n^+)] [\exp(-iW_{n'}^+) \vec{\nabla} \exp(iW_n^+)] dR \}, \end{aligned} \quad (9.9)$$

where the  $b$ 's are the coefficients of the expansion of  $\Psi_b$ :

$$\Psi_b = \sum_n b_n \Psi_n. \quad (9.10)$$

Having dropped the parts of the WKB functions which refer to irrelevant branches, we may now extend the contour  $\Sigma_{r,r}$  until its ends coincide with points at which maximum destructive interference (of the  $W^+$ ,  $W^+$  parts) occurs [e.g., the points  $A$  and  $B$  in Fig. 1(c)]. The contour then corresponds to a proper time lapse of  $T/2\Delta n$  instead of  $T/4\Delta n$ , and spans just the right number of nodes so that the integral in (9.9) vanishes except when  $n=n'$ . Expression (9.9) accordingly reduces

to

$$\begin{aligned} (\Psi_b, \Psi_a) &= \sum_n b_n * a_n \int_{\Sigma} [(\omega_n \Delta n / 2\pi |V_n|) dq \\ &\quad - (\omega_n \Delta n / 2\pi |V_{n'}|) dR] \\ &= \sum_n b_n * a_n [(\omega_n T / 4\pi) + (\omega_n T / 4\pi)], \end{aligned} \quad (9.11)$$

the positive sign of the final bracketed factor being obtained by appropriately orienting the original contour  $\Sigma_{r,r}$ . The contour  $\Sigma_{1,2}$  in Fig. 1(c) shows the correct orientation. If the direction of "motion" of the packets

is reversed then the orientation of the contour must be reversed.

For good packets the frequencies  $\omega_n$  and  $\omega_{n'}$  remain sensibly constant and equal to the peak frequencies  $2\pi/T$  and  $2\pi/\tau$ , respectively, over the range of effective  $n$  values in the sum (9.11). Therefore we have

$$(\Psi_b, \Psi_a) = \sum_n b_n^* a_n, \quad (9.12)$$

which is just the accepted definition. By virtue of the  $\Sigma$  invariance of expression (9.8) the contour may now be displaced to any location, including turning points where the WKB approximation breaks down. All that is required is that the contour cut each wave packet only once and that  $\Psi_a$  and  $\Psi_b$  vanish at its endpoints. We therefore have quite generally

$$\int_{\Sigma} (\Psi_b^* \vec{V} \Psi_a dq - \Psi_b^* \vec{V} \Psi_a dR) \approx \sum_n b_n^* a_n, \quad (9.13)$$

the relation “ $\approx$ ” tending toward “ $=$ ” the more precisely defined the packets  $\Psi_a$  and  $\Psi_b$  become.

If, in the above derivation, the two packets had been oppositely oriented then one of the pairs of functions  $\hat{W}^+$ ,  $\hat{W}^+$  in Eq. (9.9) would have had to be changed to  $\hat{W}^-$ ,  $\hat{W}^-$ , and the integral, with the extended contour, would have vanished even when  $n=n'$ . This, however, does not conflict with (9.12) since, in the case of oppositely oriented packets, the relative phases of  $a_n$  and  $b_n$  vary so rapidly with  $n$  that the inner product vanishes anyway.

An entirely similar analysis can be carried out in the  $R-\tau$  and  $\tau-q$  planes. Here the inner product integrals are given by

$$(\Phi_b, \Phi_a) = \int_{\Sigma_{\tau, t}} (\Phi_b^* \vec{V} \Phi_a d\tau - \Phi_b^* \Phi_a dR), \quad (9.14)$$

$$(\Phi_a, \Phi_b) = \int_{\Sigma_{t, \tau}} (\Phi_b^* \Phi_a dq - \Phi_b^* \vec{V} \Phi_a d\tau), \quad (9.15)$$

which reduce to the familiar  $\int \Phi_b^* \Phi_a dR$  and  $\int \Phi_b^* \Phi_a dq$  when the contours are distorted to  $\tau=\text{constant}$  and  $\tau=\text{constant}$ , respectively. If the ranges of integration of the latter integrals are extended to include all permissible  $R$  and  $q$  values, the integrals must be divided by  $\Delta n$  and  $\Delta n$ , respectively, because of the presence of the ghost packets.

The above analysis permits us to adopt a new viewpoint regarding the Cauchy problem for the Hamiltonian constraint. At the end of Sec. 6 it was conjectured that by virtue of the boundary condition (6.31) [(7.15) in the present context] the state function will be determined everywhere as soon as it is specified on a hypersurface. This is very easy to demonstrate in the present context, because of the separability of Eq. (7.12), which permits the eigenfunctions  $\Psi_n$  to be expressed in the product form

$$\Psi_n(R, q) = (2\pi\Delta n/\omega_n)^{1/2} X_n(R) X_n(q), \quad (9.16)$$

where  $X_n$  and  $X_{n'}$  have the WKB approximations

$$X_n = (\omega_n/2\pi|V_n|)^{1/2} \times [\exp(-iW_n^+) + \exp(-iW_n^-)], \quad (9.17)$$

$$X_{n'} = (\omega_{n'}/2\pi|V_{n'}|)^{1/2} \times [\exp(iW_{n'}^+) + \exp(iW_{n'}^-)], \quad (9.18)$$

and satisfy the orthonormality conditions

$$\int_0^\infty X_n^* X_{n'} dR = \delta_{nn'}, \quad \int X_n^* X_{n'} dq = \delta_{nn'}. \quad (9.19)$$

Thus, we may write

$$\Psi(R, q) = \sum_n X_n(R) X_n(q) \int X_n^*(q') \times \Psi(R', q') dq'/X_n(R') \quad (9.20a)$$

$$= \sum_n X_n(R) X_n(q) \int_0^\infty X_n^*(R') \times \Psi(R', q') dR'/X_n(q'), \quad (9.20b)$$

which express  $\Psi$  everywhere in terms of its values on the infinite contour  $R=R'$  or on the infinite contour  $q=q'$ .

However, when the function  $\Psi$  has the form of a wave packet  $\Psi_a$ , it should be equally possible to determine it completely by knowing its value over a *finite* contour  $\Sigma$  which intersects the packet only once. That this is indeed the case follows from the fact that for a good packet the integrals

$$\int_{\Sigma} (\Psi_n^* \vec{V} \Psi_a dq - \Psi_n^* \vec{V} \Psi_a dR), \quad (9.21)$$

for all  $n$ , may to a high degree of accuracy be replaced simply by

$$\int_{\Sigma} (\Psi_n^* |V_n| \Psi_a dq - \Psi_n^* |V_n| \Psi_a dR). \quad (9.22)$$

When the packet is good these integrals have non-negligible values only for a restricted range of  $n$  values centered on the peak of the packet, over which  $|V_n|$  varies slowly. Let the peak  $n$  values be determined by evaluating (9.22) for all  $n$ . Then let the contour be extended until its ends reach points of maximum destructive interference, as determined by the peak  $n$  value and the slope of the packet branch where it intersects  $\Sigma$ .<sup>49</sup> Suppose the slope is NE-SW, corresponding to the classical functions  $W^+$ ,  $W^+$  or  $W^-$ ,  $W^-$ . Denote by

<sup>49</sup> It may be objected that in choosing  $\Sigma$  to intersect the packet along a definite branch we are assuming some preliminary knowledge about the approximate “location” of the packet. This preliminary knowledge, however, differs in no fundamental respect from the knowledge which we always have in other more familiar instances, e.g., that a given particle is “somewhere in the laboratory.”

$\Psi_n^{++}$  and  $\Psi_n^{--}$  the parts of (the WKB approximation to)  $\Psi_n$  associated with these functions. Now compute the integrals

$$\int_{\Sigma'} (\Psi_n^{\pm\pm*} | V_n | \Psi_n d\eta - \Psi_n^{\pm\pm*} | V_n | \Psi_n dR), \quad (9.23)$$

where  $\Sigma'$  denotes the extended contour. Of these integrals, only those corresponding to the previously determined significant  $n$  values need be included, and of these, in turn, only those corresponding to a definite choice of signs (either  $++$  or  $--$ ) will have non-negligible values (corresponding to a definite packet orientation, which becomes thus determined). The values in question are just the amplitudes  $a_n$  of Eq. (8.1) and from these the entire state function can be constructed. This means that for a good packet the Cauchy data are not only the same as for the ordinary Schrödinger equation but are also effectively taken from a compact domain of "configuration space."

## 10. DISCUSSION AND SPECULATION

Perhaps the most impressive fact which emerges from a study of the quantum theory of gravity is that it is an extraordinarily economical theory. It gives one just exactly what is needed in order to analyze a particular physical situation, but not a bit more. Thus it will say nothing about time unless a clock to measure time is provided, and it will say nothing about geometry unless a device (either a material object, gravitational waves, or some other form of radiation) is introduced to tell when and where the geometry is to be measured.<sup>50</sup> In view of the strongly operational foundations of both the quantum theory and general relativity this is to be expected. When the two theories are united the result is an operational theory *par excellence*.<sup>51</sup>

The economy of quantum gravodynamics is also revealed in the manner in which the formalism determines its own interpretation. We have seen how the Hamiltonian constraint, in the case of a finite universe, forces us to abandon all use of externally imposed coordinates (in particular  $x^0$ ) and to look instead for an internal description of the dynamics. We have seen

<sup>50</sup> For details on the quantum theory of measurement in general relativity see B. S. DeWitt, in *Gravitation: An Introduction to Current Research*, edited by L. Witten (John Wiley & Sons, Inc., New York, 1962).

<sup>51</sup> A notable failure to recognize this fact is to be found in P. W. Bridgman, in *Albert Einstein, Philosopher-Scientist*, edited by P. A. Schilpp (Tudor Publishing Company, New York, 1949). Bridgman's confusion, which is shared by others, stems from the fact that in traditional formulations of general relativity one speaks about things, such as curvilinear coordinates, which have no operationally defined reality. This confusion would have been eliminated had modern coordinate-independent formulations of differential geometry been available in 1916. Modern methods make it plain that coordinate systems are precisely what general relativity is *not* talking about. General relativity is concerned with those attributes of physical reality which are coordinate-independent and is the rock on which present day emphasis on invariance principles will ultimately stand or fall.

how the metric structure of the manifold  $\mathcal{M}$ , with its frontier of infinite curvature, suggests a natural boundary condition for the state functional, which may simplify the Cauchy data needed to specify a state. And finally, if it be permitted to extend the results of our study of the Friedmann model to the general case, we have learned how (and when) to use the inner-product definition (5.19), by recognizing that probability flows in closed circuits in  $\mathcal{M}$ .

This "principle of self-determination," which permeates even classical general relativity, has been elevated to the rank of a universal principle by Everett,<sup>52</sup> who applies it to ordinary nonrelativistic quantum mechanics. As conventionally formulated quantum mechanics comes in two packages: (1) formalism and (2) interpretation based on the existence of a classical level. According to Everett, package 2 should be thrown away. Quantum mechanics is a theory which attempts to describe in mathematical language a situation in which chance is not a measure of our ignorance but is *absolute*. Naturally it cannot avoid introducing things like wave functions which undergo repeated fission, corresponding to the many possible outcomes of a given physical process. According to Everett, the wave function nonetheless provides a faithful representation of reality; it is the universe itself which splits.

To those who would immediately object that they do not feel themselves split, Everett replies that this only confirms the theory; they are not supposed to feel it. Everett allows into the theory only those elements which are in the formalism itself, namely, a Hilbert space, a Hamiltonian, and a Schrödinger equation for vectors in the Hilbert space. From these meager beginnings one can show, by standard arguments, that the wave function for a Hamiltonian which, in conventional language, would be described as that of a system coupled to an apparatus, evolves into a superposition of vectors representing the possible values of some system variable together with corresponding apparatus "readings." Moreover, if the "measurement" is repeated on a large number  $N$  of identically prepared systems, the final superposition consists of vectors representing various possible sets of  $N$  values for the system variable together with corresponding apparatus "memory sequences" which record these values. No interpretation of the mathematics is admitted up to this point; in particular no *a priori* interpretation is given to the coefficients in the final superposition.

Now let the coefficients in the final superposition in the case of a single system be denoted by  $c_n$ . Then the coefficients in the case of  $N$  systems will be products of  $c$ 's. It can be shown<sup>53</sup> that if one removes from the final  $N$ -system superposition all those vectors which correspond to memory sequences in which the recorded values of the system variable fail to meet the standard

<sup>52</sup> H. Everett, III, Rev. Mod. Phys. **29**, 454 (1957).

<sup>53</sup> N. R. Graham, Ph.D. thesis, University of North Carolina (unpublished).

requirements for a *random sequence with probabilities*  $|c_n|^2$ , to any arbitrary, but fixed, degree of accuracy, the resulting wave function is indistinguishable from the true final wave function in the limit  $N \rightarrow \infty$ . By "indistinguishable" we mean that the difference between it and the true wave function has vanishing norm.

The probability interpretation of quantum mechanics thus emerges from the formalism itself. Nonrandom memory sequences are "of measure zero" in the final superposition, in the limit  $N \rightarrow \infty$ . Each automaton (i.e., apparatus *cum* memory sequence) in the superposition sees the world obey the familiar quantum laws. However, there exists no outside agency which can designate which "branch" of the superposition is to be regarded as the *real* world. All are equally real, and yet each is unaware of the others. Thus if, within a given branch, an automaton, which has measured a given variable without changing it, subsequently checks his original observation, his memory sequence will not fail him. He will get his original value, and not that of some other branch. Moreover, if he communicates with another automaton who has simultaneously made the same measurement, their results will agree, which means that the two are in the same branch and that communication between different branches is impossible. The automaton therefore never feels himself split.

Everett's view of the world is a very natural one to adopt in the quantum theory of gravity, where one is accustomed to speak without embarrassment of the "wave function of the universe." It is possible that Everett's view is not only natural but essential. For example, if the Hamiltonian constraint possesses only a single solution, so that the wave function for the universe is unique, then some conception like Everett's would appear to be needed in order to assess the physical significance of such uniqueness.<sup>54</sup>

In our discussion of the Friedmann model we assumed that the operator  $\mathcal{H} + \mathcal{J}$  possesses a highly degenerate zero eigenvalue. How plausible is this assumption in the case of the actual world? In the case of the Friedmann model we were obliged to match the internal dynamics of the dust with that of the universe as a whole, with one hundred percent precision. Let us try to be a little more realistic. Suppose we replace the dust by a gas of noninteracting scalar bosons, but still maintain a rigid spherical geometry. Then we have an infinity of degrees of freedom. However, this infinity is *discrete*, because the universe is finite. Moreover, and this is important, there can never be more than a finite number of field quanta present in the state vector superposition, since the total energy (of the bosons) cannot be infinite. This is true even if the bosons are massless, since there is no infrared catastrophe in a finite world.

Now it is not at all difficult to verify that the Hamiltonian  $\mathcal{H}$  in this case does not match  $\mathcal{J}$  in any obviously

commensurable way.<sup>55</sup> For each choice of boson quantum numbers  $\mathcal{J}$  becomes a well-defined function of  $R$ , and the combination  $\mathcal{H} + \mathcal{J}$  has a well-defined spectrum. But only by the sheerest accident does this spectrum include zero. All the evidence points to the fact that the complete spectrum of  $\mathcal{H} + \mathcal{J}$ , although discrete, is everywhere dense on the real line and does not condense into a set of finitely separated, infinitely degenerate levels. A similar situation holds with vector bosons and with fermions, and it seems hardly likely that the switching on of interactions between the particles will change the picture.

One might now suggest that we look for a way out of this predicament by relaxing the spherical rigidity restriction. However, this would merely correspond to the introduction of a gas of interacting tensor bosons, i.e., gravitons. It therefore appears that the same situation holds even in the general theory and that the Hamiltonian constraint of the real world may indeed have only one solution.<sup>56</sup>

If the state functional of the universe is unique how can we interpret it? In the case of the Friedmann model a single eigenfunction  $\Psi_n$  certainly has no resemblance to the real world, nor to any other reasonable world for that matter. A plot of  $|\Psi_n|^2$  in the  $R-q$  plane looks like a lot of bumps separated from one another by a rectangular array of nodal lines, certainly nothing like a Lissajous figure. However, suppose an extra term were added to the Hamiltonian  $\mathcal{H} + \mathcal{J}$  which had the effect of strongly correlating the phase of the coherent particle clock with the phase of the universe, without changing either the (zero) eigenvalue or the quantum numbers. Then the Hamiltonian would no longer be separable and the nodal lines of  $|\Psi_n|^2$  would no longer form a rectangular array. The bumps would instead tend to cluster around the Lissajous figure having the favored correlation, the figure itself now being somewhat distorted due to the correlation interaction, but still definitely recognizable.

The Hamiltonian of the real world is highly nonseparable, and there is a high degree of correlation among its infinity of modes. This must express itself as a kind of "condensation" of the state functional into components having many of the attributes of the quasiclassical Friedmann packets.<sup>57</sup> At the same time, because of the size of the universe, we know that the "Everett process" must be occurring on a lavish scale: The quasiclassical components of the universal state

<sup>55</sup> M. Miketinac (private communication).

<sup>56</sup> The spectrum of  $\mathcal{H} + \mathcal{J}$ , or of  $\mathcal{H}$  itself, can be shifted by the introduction of a "cosmological term" in the Einstein Lagrangian. If this spectrum is actually everywhere dense then we have the amusing result that a minute change in the cosmological constant can produce an enormous change in the zero-eigenvalue eigenvector and hence in the physical properties of the universe.

<sup>57</sup> If the state functional of the universe is unique then it is no longer possible or even meaningful to apply the inner-product definition (5.19) to the state functional as a whole. However, it might still be applied, in some reduced form, to its quasiclassical components.

<sup>54</sup> See J. A. Wheeler, *The Monist* 47, 40 (1962).

functional must be constantly splitting into a stupendous number of branches, all moving in parallel without interfering with one another except insofar as quantum Poincaré cycles allow rare anomalies to occur. According to the Everett interpretation each branch corresponds to a possible world-as-we-actually-see-it.

We have seen that the Friedmann packets in the  $R-q$  plane do not ultimately spread in "time"; every expansion-contraction cycle is exactly like every other. Unless some form of leakage to other channels occurs (e.g., transitions to different 3-space topologies) the same must be true for the real universe (assuming it to be closed and finite). In the absence of such channels there could be only *one* expansion-contraction cycle, repeated over and over again, like a movie film, throughout eternity, the monotony of which would be alleviated only by the infinite variety to be found among the multitude of simultaneous parallel worlds all executing the cycle together. Such a conclusion holds, in fact, regardless of whether the total state functional is unique or not.

A question naturally arises in regard to entropy. Within a given branch of the universal state functional the entropy would be observed (by appropriate automata) to increase with time.<sup>58</sup> It might be supposed that this increase would continue only during the expansion phase of the universe and that it would reverse itself during the contraction phase. This is not so, for one has only to remember that the length of a Poincaré cycle for even a small part of the universe is vastly longer than a rebound cycle, and hence except for a vanishingly small fraction of branches the entropy must continue to increase (at least locally) until final collapse is reached, at which point the very concepts of entropy and probability, as well as time itself, cease to have meaning.

However, if the operator  $\mathcal{H} + \mathcal{C}$  is time-reversal invariant, and if its zero eigenvalue is nondegenerate, then the state functional of the universe is necessarily time-symmetric. This means that for every Everett branch in which entropy increases with time there must be another in which entropy decreases with time. To an observer in the second branch "time" in fact appears to be "flowing" in the opposite sense. Because of the extreme sensitivity of the state functional to slight changes in the operator  $\mathcal{H} + \mathcal{C}$  (see Ref. 56) it is difficult to say how these conclusions must be modified if, as recent experiments suggest, the real world is not invariant under time reversal. However, the world being as

<sup>58</sup> Each branch corresponds to a pure state in the traditional sense. This does not, however, prevent the assignment of an effective entropy to it. For a sufficiently complicated system even a pure state may be assigned an entropy based on the coarse-grained properties of the state rather than on an ensemble average. In the classical theory this is illustrated by computer calculations of  $n$ -body systems. Even though the position and velocity of every body is known, the system as a whole possesses effective thermodynamical properties, the determination of which is in fact the goal of the computation.

complicated (and hence ergodic) as it is, it is still quite possible that there is no *preferred* direction in time. The ensemble of Everett branches in which time has a given direction of flow may very well be balanced by another ensemble in which time flows oppositely, so that reality as a whole possesses no over-all time orientation despite the absence of time-reversal invariance.

#### ACKNOWLEDGMENTS

I wish to express my warmest gratitude for the kindness of Professor Robert Oppenheimer and Professor Carl Kaysen in extending to me the hospitality of the Institute for Advanced Study.

#### APPENDIX A: THE MANIFOLD $M$

$M$  is defined as the 6-dimensional space of "points"  $\{\gamma_{ij}\}$  having as covariant and contravariant metric tensors, respectively, the expressions

$$G^{ijkl} = \frac{1}{2} \gamma^{1/2} (\gamma^{ik}\gamma^{jl} + \gamma^{il}\gamma^{jk} - 2\gamma^{ij}\gamma^{kl}), \quad (A1)$$

$$G_{ijkl} = \frac{1}{2} \gamma^{-1/2} (\gamma_{ik}\gamma_{jl} + \gamma_{il}\gamma_{jk} - \gamma_{ij}\gamma_{kl}), \quad (A2)$$

satisfying

$$G_{ijab}G^{abkl} = \delta_{ij}{}^{kl}. \quad (A3)$$

Index pairs may, if desired, be mapped into single indices according to the rules

$$\begin{aligned} \gamma_{11} &= \gamma^1, & \gamma_{22} &= \gamma^2, & \gamma_{33} &= \gamma^3, \\ \gamma_{23} &= 2^{-1/2}\gamma^4, & \gamma_{31} &= 2^{-1/2}\gamma^5, & \gamma_{12} &= 2^{-1/2}\gamma^6, \\ \delta_{ij}{}^{kl} &\rightarrow \delta^{\Gamma}_{\Delta}, & \Gamma, \Delta &= 1 \cdots 6, \text{ etc.} \end{aligned} \quad (A4)$$

although this is seldom convenient or necessary.

By straightforward computation one may verify the variational law

$$G_{ijkl}\delta G^{ijkl} = -\gamma^{ij}\delta\gamma_{ij}, \quad (A5)$$

from which it may be inferred that

$$G \equiv \det(G^{ijkl}) = -a\gamma^{-1}, \quad (A6)$$

where  $a$  is some constant. In the special case  $\gamma_{ij} = \delta_{ij}$  one easily finds that the roots of  $G^{ijkl}$  are  $-\frac{1}{2}, 1, 1, 1, 1, 1$ , from which it follows that  $a = \frac{1}{2}$  and that the signature of  $M$  is  $-+++++$ . The components  $\gamma_{ij}$  evidently are "good" coordinates in  $M$  as long as  $\gamma \neq 0$ .

If the  $\xi$  of Eq. (5.7) is chosen as a new coordinate then the surfaces of constant  $\xi$  have orthogonal trajectories whose tangent vectors are proportional to

$$\partial\xi/\partial\gamma_{ij} = \frac{1}{4}\xi\gamma^{ij}, \quad (A7)$$

or, in contravariant form,

$$G_{ijkl}\partial\xi/\partial\gamma_{kl} = -\frac{4}{3}\xi^{-1}\gamma_{ij}. \quad (A8)$$

If the orthogonal trajectories themselves are labeled by a set of five additional new coordinates  $\xi^A$ ,  $A = 1 \cdots 5$ , then

$$\gamma^{ij}\partial\gamma_{ij}/\partial\xi^A = 4\xi^{-1}\partial\xi/\partial\xi^A = 0, \quad (A9)$$

and, moreover,  $\partial\gamma_{ij}/\partial\xi$  must satisfy

$$G^{ijkl} \frac{\partial\gamma_{ij}}{\partial\xi} \frac{\partial\gamma_{kl}}{\partial\xi^A} = 0. \quad (\text{A10})$$

From these facts one may infer

$$\frac{\partial\gamma_{ij}}{\partial\xi} \frac{\partial\gamma_{kl}}{\partial\xi} = \left( G_{ijkl} \frac{\partial\xi}{\partial\gamma_{ij}} \frac{\partial\xi}{\partial\gamma_{kl}} \right)^{-1} = -1, \quad (\text{A11})$$

$$\partial\gamma_{ij}/\partial\xi = \frac{4}{3}\xi^{-1}\gamma_{ij}, \quad (\text{A12})$$

$$\gamma_{ij}\partial\xi^A/\partial\gamma_{ij} = 0, \quad (\text{A13})$$

$$G^{ijkl} \frac{\partial\gamma_{ij}}{\partial\xi^B} \frac{\partial\gamma_{kl}}{\partial\xi^B} = \gamma^{1/2} \gamma^{ik} \gamma^{jl} \frac{\partial\gamma_{ij}}{\partial\xi^A} \frac{\partial\gamma_{kl}}{\partial\xi^B}, \quad (\text{A14})$$

from which the metric (5.8) follows.

The above relations yield the following useful identities:

$$\text{tr}(\gamma_{,A}\partial\xi^B/\partial\gamma) = \delta_A^B, \quad (\text{A15})$$

$$\text{tr}(\gamma\partial\xi^A/\partial\gamma) = 0, \quad (\text{A16})$$

$$\text{tr}(\gamma^{-1}\gamma_{,A}) = 0, \quad (\text{A17})$$

$$\text{tr}[\gamma^{-1}(\gamma_{,AB} - \gamma_{,A}\gamma^{-1}\gamma_{,B})] = 0, \quad (\text{A18})$$

$$\frac{\partial\gamma_{ij}}{\partial\xi^A} \frac{\partial\xi^A}{\partial\gamma_{kl}} = \delta_{ij}^{kl} - \frac{1}{3}\gamma_{ij}\gamma^{kl}, \quad (\text{A19})$$

$$\begin{aligned} &\text{tr}(\gamma_{,A}\mathbf{M})\text{tr}(\mathbf{N}\partial\xi^A/\partial\gamma) \\ &= \frac{1}{2}\text{tr}(\mathbf{M}\mathbf{N} + \mathbf{M}\mathbf{N}^\sim) - \frac{1}{3}\text{tr}(\gamma\mathbf{M})\text{tr}(\gamma^{-1}\mathbf{N}), \end{aligned} \quad (\text{A20})$$

$$\begin{aligned} &\text{tr}(\gamma_{,AB}\mathbf{M})\text{tr}(\mathbf{N}\partial\xi^B/\partial\gamma) + \text{tr}(\gamma_{,B}\mathbf{M})\text{tr}[\mathbf{N}(\partial\xi^B/\partial\gamma)_{,A}] \\ &= -\frac{1}{3}\text{tr}(\gamma_{,A}\mathbf{M})\text{tr}(\gamma^{-1}\mathbf{N}) \\ &\quad + \frac{1}{3}\text{tr}(\gamma\mathbf{M})\text{tr}(\gamma^{-1}\gamma_{,A}\gamma^{-1}\mathbf{N}), \end{aligned} \quad (\text{A21})$$

$$\begin{aligned} &\text{tr}[\gamma_{,A}\mathbf{M}(\partial\xi^A/\partial\gamma)\mathbf{N}] \\ &= \frac{1}{2}\text{tr}(\mathbf{M}\mathbf{N}^\sim) + \frac{1}{2}\text{tr}\mathbf{M}\text{tr}\mathbf{N} - \frac{1}{3}\text{tr}(\gamma\mathbf{M})\gamma^{-1}\mathbf{N}). \end{aligned} \quad (\text{A22})$$

Equations (A20) and (A22) follow from (A19), while Eqs. (A18) and (A21) are obtained from (A17) and (A20), respectively, by differentiation.  $\mathbf{M}$  and  $\mathbf{N}$  are arbitrary  $3 \times 3$  matrices.

Using these identities and remembering the cyclic invariance of the trace, it is easy to compute the following:

$$\begin{aligned} \bar{\Gamma}^{AB} &= \text{tr}[\gamma(\partial\xi^A/\partial\gamma)\gamma(\partial\xi^B/\partial\gamma)], \\ \bar{G}_{AC}\bar{G}^{CB} &= \delta_A^B, \end{aligned} \quad (\text{A23})$$

$$\begin{aligned} \bar{\Gamma}_{ABC} &\equiv \frac{1}{2}(\bar{G}_{AC,B} + \bar{G}_{BC,A} - \bar{G}_{AB,C}) = \frac{1}{2}\text{tr}[\gamma_c\gamma^{-1} \\ &\quad \times (-\gamma_{,A}\gamma^{-1}\gamma_{,B} - \gamma_{,B}\gamma^{-1}\gamma_{,A} + 2\gamma_{,AB})\gamma^{-1}], \end{aligned} \quad (\text{A24})$$

$$\begin{aligned} \bar{\Gamma}_{AB}^C &\equiv \bar{G}^{CD}\bar{\Gamma}_{ABD} \\ &= \text{tr}[(-\gamma_{,A}\gamma^{-1}\gamma_{,B} + \gamma_{,AB})\partial\xi^C/\partial\gamma], \end{aligned} \quad (\text{A25})$$

$$\begin{aligned} \bar{R}_{ABC}^D &\equiv \bar{\Gamma}_{BC}^D,_A - \bar{\Gamma}_{AC}^D,_B + \bar{\Gamma}_{BC}^E\bar{\Gamma}_{AE}^D - \bar{\Gamma}_{AC}^E\bar{\Gamma}_{BE}^D \\ &= \text{tr}[\gamma_c\gamma^{-1}(\gamma_{,A}\gamma^{-1}\gamma_{,B} - \gamma_{,B}\gamma^{-1}\gamma_{,A}) \\ &\quad \times \partial\xi^D/\partial\gamma], \end{aligned} \quad (\text{A26})$$

$$\begin{aligned} \bar{R}_{ABCD} &\equiv \bar{R}_{ABC}^E\bar{G}_{ED} = \text{tr}[\gamma^{-1}\gamma_{,D}\gamma^{-1}\gamma_c\gamma^{-1} \\ &\quad \times (\gamma_{,A}\gamma^{-1}\gamma_{,B} - \gamma_{,B}\gamma^{-1}\gamma_{,A})], \end{aligned} \quad (\text{A27})$$

$$\bar{R}_{AB} \equiv \bar{R}_{CAB}^C = -\frac{3}{2}\bar{G}_{AB}. \quad (\text{A28})$$

The corresponding quantities in the full manifold  $M$  are

$$\Gamma_{AB}^C = \bar{\Gamma}_{AB}^C, \quad (\text{A29})$$

$$\Gamma_{AB}^0 = (3/32)\xi\bar{G}_{AB}, \quad (\text{A30})$$

$$\Gamma_{A0}^B = \xi^{-1}\delta_A^B, \quad (\text{A31})$$

$$\Gamma_{A0}^0 = \Gamma_{00}^A = \Gamma_{00}^0 = 0, \quad (\text{A32})$$

$$\begin{aligned} R_{ABC}^D &= \bar{R}_{ABC}^D - (3/32)(\bar{G}_{AC}\delta_B^D - \bar{G}_{BC}\delta_A^D), \\ &\quad + (\text{all other components vanish}) \end{aligned} \quad (\text{A33})$$

$$R_{AB} = \bar{R}_{AB} + \frac{3}{8}\bar{G}_{AB} = -(9/8)G_{AB}, \quad (\text{A34})$$

$$R_{A0} = R_{00} = 0, \quad (\text{A35})$$

$${}^{(6)}R = -60\xi^{-2}, \quad (\text{A36})$$

where the index 0 is used for components in the direction of the “timelike” coordinate  $\xi$ .

The geodesic equation in  $\bar{M}$  is obtained directly from (A25):

$$\begin{aligned} 0 &= \frac{d^2\xi^A}{ds^2} + \bar{\Gamma}_{BC}^A \frac{d\xi^B}{ds} \frac{d\xi^C}{ds} \\ &= \text{tr}\left[\frac{\partial\xi^A}{\partial\gamma} \left(\frac{d^2\gamma}{ds^2} - \frac{d\gamma}{ds}\gamma^{-1}\frac{d\gamma}{ds}\right)\right]. \end{aligned} \quad (\text{A37})$$

This reduces to (5.11) upon multiplication with  $\gamma_{,A}$  and use of (A17), (A18), and (A19).

When one stays within the manifold  $\bar{M}$  it is convenient to map the matrices  $\gamma$  into matrices  $a$  of unit determinant:

$$a \equiv \gamma^{-1/3}\gamma. \quad (\text{A38})$$

In the solution (5.12) of the geodesic equation,  $\gamma$  may be replaced by  $a$  provided  $\mathbf{M}$  is restricted to have unit determinant. To find the geodesic connecting two matrices  $\mathbf{a}_1$  and  $\mathbf{a}_2$ , let  $\bar{s}=0$  at  $\mathbf{a}_1$  and choose  $\mathbf{M}$  in the form

$$\mathbf{M} = \mathbf{d}_1^{1/2}\mathbf{O}, \quad (\text{A39})$$

where  $\mathbf{O}$  is an orthogonal matrix which diagonalizes  $\mathbf{a}_1$  and  $\mathbf{d}_1^{1/2}$  is a diagonal square root of the resulting diagonal matrix. Then if  $\bar{s}_{12}$  is the distance between  $\mathbf{a}_1$  and  $\mathbf{a}_2$  the matrix  $\mathbf{N}$  satisfies

$$\bar{s}_{12} = \ln(\mathbf{d}_1^{-1/2}\mathbf{O}\mathbf{a}_2\mathbf{O}^\sim\mathbf{d}_1^{-1/2}). \quad (\text{A40})$$

From the condition  $\text{tr}\mathbf{N}^2 = 1$  one obtains

$$\begin{aligned} \bar{s}_{12}^2 &= \text{tr}[\ln(\mathbf{d}_1^{-1/2}\mathbf{O}\mathbf{a}_2\mathbf{O}^\sim\mathbf{d}_1^{-1/2})]^2 \\ &= \text{tr}[\ln(\mathbf{a}_1^{-1}\mathbf{a}_2)]^2, \end{aligned} \quad (\text{A41})$$

which permits the matrix  $\mathbf{N}$  itself to be determined. The logarithm of a matrix is an effectively unambiguous

concept, and the law of cyclic invariance of the trace applies to transcendental matrix functions as well as to rational functions. The uniqueness of the geodesic (5.12) is easily checked by noting that the matrix  $\mathbf{O}$  is determined up to a transformation of the form  $\mathbf{O}'=\mathbf{P}\mathbf{O}$  where  $\mathbf{P}$  is either a permutation matrix, if the roots of  $\mathbf{a}_1$  are all distinct, or a more general orthogonal matrix, if some of the roots coincide. Such a transformation leaves (5.12) invariant. It is also easy to verify that it does not matter which of the eight possible square roots of  $\mathbf{d}_1$  is chosen for  $\mathbf{d}_1^{1/2}$ .

The geodesic equations in  $M$  take the form

$$0 = \frac{d^2\xi}{ds^2} + (3/32)\xi \left( \frac{d\xi}{ds} \right)^2, \quad (\text{A42})$$

$$0 = \frac{d^2\xi^A}{ds^2} + \bar{\Gamma}_{BC}^A \frac{d\xi^B}{ds} \frac{d\xi^C}{ds} + 2\xi^{-1} \frac{d\xi^A}{ds} \frac{d\xi}{ds}, \quad (\text{A43})$$

where

$$\left( \frac{d\xi}{ds} \right)^2 = \tilde{G}_{AB} \frac{d\xi^A}{ds} \frac{d\xi^B}{ds}. \quad (\text{A44})$$

Differentiating the latter equation and making use of (A43) multiplied by  $\tilde{G}_{AB} d\xi^B/ds$ , one finds

$$\frac{d^2\tilde{s}}{ds^2} = -\frac{2}{\xi} \frac{d\xi}{ds} \frac{d\xi}{ds}, \quad (\text{A45})$$

which may be integrated to yield

$$\frac{d\tilde{s}}{ds} = \frac{\alpha}{\xi^2}. \quad (\text{A46})$$

$\alpha$  is an arbitrary integration constant which, without loss of generality, may be taken positive. When  $\alpha \neq 0$  one may write

$$\frac{d\xi^A}{d\tilde{s}} = \frac{\xi^2}{\alpha} \frac{d\xi^A}{ds}, \quad (\text{A47})$$

which, in combination with (A43), yields (A37), showing that geodesics in  $M$  project onto geodesics in  $\tilde{M}$ .

Substitution of (A46) into (A42) yields

$$\frac{d^2\xi}{ds^2} + \frac{\kappa^2\alpha^2}{\xi^3} = 0, \quad \kappa = (3/32)^{1/2}, \quad (\text{A48})$$

which integrates to

$$d\xi/ds = \pm(\kappa^2\alpha^2\xi^{-2} - \beta)^{1/2}, \quad (\text{A49})$$

where  $\beta$  is another integration constant. Using the metric (5.8) it is easy to verify that standard normalization for the affine parameter  $s$  is obtained by choosing  $\beta = -1, 0$ , or  $1$  according as the geodesic is timelike, null, or spacelike.

*Timelike geodesics.* In this case  $\xi$  must always increase (or decrease) with  $s$ . Therefore choosing the positive root in (A49) and the boundary conditions  $\xi(0)=0$ ,  $\xi(\infty)=0$ , one finds, upon setting  $\beta=-1$  and

integrating Eqs. (A46) and (A49),

$$\xi(s) = [s(2\kappa\alpha + s)]^{1/2} = -\kappa\alpha \operatorname{csch}(\kappa\bar{s}), \quad (\text{A50})$$

$$\bar{s}(s) = \frac{1}{2\kappa} \ln \frac{s}{2\kappa\alpha + s}. \quad (\text{A51})$$

The ranges of the variables are

$$0 \leq s < \infty, \quad -\infty \leq \bar{s} < 0, \quad 0 \leq \xi < \infty, \quad (\text{A52})$$

and one sees that the geodesic strikes the frontier at  $s=0$ .

In terms of the matrix  $\gamma$  the above results may be expressed in the form

$$\gamma(s) = [-\kappa^2\alpha \operatorname{csch}(\kappa\bar{s})]^{4/3} \mathbf{M} \sim e^{\mathbf{N}\bar{s}} \mathbf{M}, \quad (\text{A53})$$

where  $\mathbf{N}$  is restricted as in (5.13) and  $\mathbf{M}$  is now required to have unit determinant.

*Null geodesics.* In this case it is the constant  $\alpha$  which serves to fix the scale of  $s$ . Setting  $2\kappa\alpha=1$ ,  $\beta=0$ , and choosing the positive root in (A49), one finds, with the boundary condition  $\xi(0)=0$ ,

$$\xi(s) = s^{1/2}. \quad (\text{A54})$$

There is no preferred zero point for the arc length  $\bar{s}$  in  $\mathbf{M}$ . Hence the following integral of Eq. (A46) may be chosen:

$$\bar{s} = (2\kappa)^{-1} \ln s. \quad (\text{A55})$$

This yields

$$\xi(s) = e^{\kappa\bar{s}}, \quad (\text{A56})$$

$$\gamma(s) = (\kappa e^{\kappa\bar{s}})^{4/3} \mathbf{M} \sim e^{\mathbf{N}\bar{s}} \mathbf{M}. \quad (\text{A57})$$

The ranges of the variables are

$$0 \leq s < \infty, \quad -\infty \leq \bar{s} < \infty, \quad 0 \leq \xi < \infty, \quad (\text{A58})$$

and the geodesic is again seen to hit the frontier.

*Spacelike geodesics.* In this case there is a turning point at  $\xi=\kappa\alpha$  and both roots in (A49) can occur. Setting  $\beta=1$ , and choosing the boundary conditions  $\xi(0)=0$ ,  $\xi(\kappa\alpha)=0$ , with  $s \geq 0$ , one finds

$$\xi(s) = [s(2\kappa\alpha - s)]^{1/2} = \kappa\alpha \operatorname{sech}(\kappa\bar{s}), \quad (\text{A59})$$

$$\bar{s}(s) = \frac{1}{2\kappa} \ln \frac{s}{2\kappa\alpha - s}, \quad (\text{A60})$$

$$\gamma(s) = [\kappa^2\alpha \operatorname{sech}(\kappa\bar{s})]^{4/3} \mathbf{M} \sim e^{\mathbf{N}\bar{s}} \mathbf{M}. \quad (\text{A61})$$

The ranges of the variables are

$$0 \leq s \leq 2\kappa\alpha, \quad -\infty \leq \bar{s} \leq \infty, \quad 0 \leq \xi \leq \kappa\alpha, \quad (\text{A62})$$

and the geodesic is seen to hit the frontier at both ends ( $s=0, 2\kappa\alpha$ ).

Using the above results it is possible to obtain an expression for the “distance” to the frontier from any point in  $M$ . If  $\gamma$  is a fixed point and  $\xi$  is a contravariant vector at  $\gamma$ , then

$$\sigma(\gamma, \xi) = \frac{16 [\operatorname{tr} \mathbf{X}^2 - \frac{1}{3}(\operatorname{tr} \mathbf{X})^2]^{1/2} + (\frac{2}{3})^{1/2} \operatorname{tr} \mathbf{X}}{3 [\operatorname{tr} \mathbf{X}^2 - \frac{1}{3}(\operatorname{tr} \mathbf{X})^2]^{1/2} - (\frac{2}{3})^{1/2} \operatorname{tr} \mathbf{X}} \gamma^{1/2}, \quad (\text{A63})$$

$$\mathbf{X} = \gamma^{-1}\xi,$$

where  $\sigma$  is one-half the square of the distance from  $\gamma$  to the frontier along the geodesic which starts from  $\gamma$  in the direction of  $\xi$ . The formula holds as long as  $\text{tr } X < (\text{tr } X^2)^{1/2}$ , so that the denominator is positive, for otherwise the geodesic escapes to infinity in the direction of  $\xi$ .  $\sigma$  takes on its minimum value,  $-(16/3)\gamma^{1/2}$ , when the geodesic is a path of pure dilation which arrives at the frontier at the point  $\gamma=0$ . The condition for this is  $\text{tr } X^2 = \frac{1}{3}(\text{tr } X)^2$ .

It should be noted that when the geodesic is not a path of pure dilation, it usually strikes the frontier at a point where some of the  $\gamma_{ij}$  become infinite, in spite of the fact that  $\gamma$  itself vanishes there. To see this, first observe that expressions (A53), (A57), and (A61) all have the limiting behavior

$$\gamma(s) \rightarrow (2\kappa^2\alpha e^{\kappa s})^{4/3} M \sim e^{Ns} M \text{ as } s \rightarrow -\infty. \quad (\text{A64})$$

Next note that in virtue of the conditions (5.13)  $N$  always has one root which is at least as negative as  $-(\frac{1}{6})^{1/2}$ . But  $4\kappa/3 = (\frac{1}{6})^{1/2}$ . Therefore the only way in which a blow up of the exponential can be avoided in (A64) is for the roots of  $N$  to be precisely  $-(\frac{1}{6})^{1/2}$ ,  $-(\frac{1}{6})^{1/2}$ ,  $(\frac{2}{3})^{1/2}$ . Without loss of generality  $N$  may be chosen diagonal. The limiting form of  $\gamma(s)$  in this special case is then

$$\gamma(0) = M \sim \begin{pmatrix} (2\kappa^2\alpha)^{4/3} & 0 & 0 \\ 0 & (2\kappa^2\alpha)^{4/3} & 0 \\ 0 & 0 & 0 \end{pmatrix} M. \quad (\text{A65})$$

Since  $M$  and  $\alpha$  are arbitrary (subject to  $\det M = 1$ ) it follows that any singular symmetric matrix having an odd number of vanishing roots can be reached by a geodesic. Matrices having two vanishing roots can be reached (in a finite distance) from nonsingular points, but only along paths which suffer infinite absolute acceleration at the frontier.

It is not difficult to obtain an expression for the geodesic distance between two matrices  $\gamma_1$  and  $\gamma_2$  in  $M$ . The geometry of  $M$  may be summed up in the compact formula

$$ds^2 = -d\xi^2 + \xi^2 d\bar{s}^2, \quad (\text{A66})$$

which follows from (5.8). With the introduction of the variables

$$t = \xi \cosh \kappa \bar{s}, \quad x = \xi \sinh \kappa \bar{s}, \quad (\text{A67})$$

this is converted to

$$ds^2 = -dt^2 + dx^2, \quad (\text{A68})$$

which is formally just the line element of 2-dimensional Minkowski space. Hence

$$\begin{aligned} \sigma(\gamma_1, \gamma_2) &\equiv \frac{1}{2}(s_{12})^2 = -\frac{1}{2}(t_1 - t_2)^2 + \frac{1}{2}(x_1 - x_2)^2 \\ &= \frac{1}{2}(2\xi_1 \xi_2 \cosh \kappa \bar{s}_{12} - \xi_1^2 - \xi_2^2) \\ &= (16/3)\{2(\gamma_1 \gamma_2)^{1/4} \cosh[(3/32)^{1/2} \bar{s}_{12}] \\ &\quad - \gamma_1^{1/2} - \gamma_2^{1/2}\}, \end{aligned} \quad (\text{A69})$$

where  $\bar{s}_{12}$  is given by (A41).

## APPENDIX B: THE MANIFOLDS $M^{\infty^3}$ AND $\mathcal{M}$

In discussing these manifolds it will be convenient to use an abbreviated notation which avoids the necessity of writing integral signs or excessive numbers of indices bearing various numbers of primes. This notation will be applicable to completely general manifolds and, in fact, will be used again in the following paper of this series in quite a different context, thus providing additional justification for its introduction here.

The functions  $\gamma_{ij}(x)$  will be replaced by the symbol  $\varphi^i$ . More precisely, the symbol  $\gamma$  is replaced by  $\varphi$ , and the quintuple  $(i, j, x^1, x^2, x^3)$  by the single index  $i$ . In general applications the  $\varphi$ 's may constitute either a finite discrete set of real numbers or, as here, a set of functions or "fields." When the index  $i$  has a continuous character, the summation convention for repeated indices will be understood to include integrations over the continuous labels for which it stands. In this Appendix no restriction will be placed on the range of values which the indices can assume.

The  $\varphi$ 's are "coordinates" in a manifold ( $M^{\infty^3}$  in the present case) on which a group acts. Group elements will be denoted by barred letters  $\bar{x}$ ,  $\bar{y}$ , etc. and their components in some coordinate system in the group space will be denoted by  $\bar{x}^\alpha$ ,  $\bar{y}^\beta$ , etc. For example, the 3-dimensional general coordinate transformation group may be coordinatized by the functions  $\bar{x}^i(x)$  which define the coordinate transformation  $x^i \rightarrow \bar{x}^i$ . In the condensed notation the quadruplet  $(i, x^1, x^2, x^3)$  gets replaced by the single index  $\alpha$ .

The multiplication table of the group defines a set of function(al)s  $F^\alpha[\bar{y}, \bar{x}]$  satisfying

$$F^\alpha[\bar{y}, \bar{x}] = (\bar{y} \bar{x})^\alpha. \quad (\text{B1})$$

For example, in the case of the coordinate transformation group this functional has the form

$$F^i[\bar{x}, \bar{x}] = \bar{y}^i(\bar{x}(x)).$$

By virtue of the group postulates  $F^\alpha[\bar{x}, \bar{y}]$  also satisfies the following fundamental identities:

$$F^\alpha[\bar{x}, e] = F^\alpha[e, \bar{x}] = \bar{x}^\alpha, \quad (\text{B2})$$

$$F^\alpha[\bar{x}, \bar{x}^{-1}] = F^\alpha[\bar{x}^{-1}, \bar{x}] = e^\alpha, \quad (\text{B3})$$

$$F^\alpha[\bar{z}, \bar{y} \bar{x}] = F^\alpha[\bar{z} \bar{y}, \bar{x}], \quad (\text{B4})$$

where  $e$  is the identity element of the group and  $\bar{x}^{-1}$  denotes the inverse of  $\bar{x}$ . In the case of the coordinate transformation group, we have  $e^i(x) = x^i$ .

Instead of dealing directly with  $F^\alpha[\bar{y}, \bar{x}]$ , one more often makes use of a set of auxiliary function(al)s, together with the structure constants of the group. These are defined, respectively, by

$$L^\alpha_\beta[\bar{x}] \equiv (\delta F^\alpha[\bar{y}, \bar{x}] / \delta \bar{y}^\beta)_{\bar{y}=e}, \quad (\text{B5})$$

$$\begin{aligned} c^\alpha_{\beta\gamma} &\equiv (\delta^2 F^\alpha[\bar{y}, \bar{x}] / \delta \bar{y}^\beta \delta \bar{x}^\gamma)_{\bar{y}=\bar{x}=e} \\ &\quad - \delta^2 F^\alpha[\bar{y}, \bar{x}] / \delta \bar{y}^\gamma \delta \bar{x}^\beta \end{aligned} \quad (\text{B6})$$

In the case of the coordinate transformation group one finds  $L^i_{j'}[\bar{x}] = \delta^i_j \delta(\bar{x}(\mathbf{x}), \mathbf{x}')$ , while the structure constants are as given in Eq. (4.17). In the case of finite-dimensional Lie groups the functional derivatives in (B5) and (B6) become ordinary derivatives.

By repeatedly differentiating Eqs. (B2), (B3), (B4) and setting various elements equal to the identity, a number of important relations can be established. Among them we cite the following:

$$L^\alpha_\beta[e] = \delta^\alpha_\beta, \quad (\text{B7})$$

$$L^{-1\alpha}_{\delta, \epsilon} - L^{-1\alpha}_{\epsilon, \delta} = -c^\alpha_{\beta\gamma} L^{-1\beta}_\delta L^{-1\gamma}_\epsilon, \quad (\text{B8})$$

$$c^\alpha_{\beta\epsilon} c^\epsilon_{\gamma\delta} + c^\alpha_{\gamma\epsilon} c^\epsilon_{\delta\beta} + c^\alpha_{\delta\epsilon} c^\epsilon_{\beta\gamma} = 0. \quad (\text{B9})$$

In Eq. (B8) the arguments of the function(al)s have been suppressed, and differentiation is denoted by a comma.  $L^{-1\alpha}_\beta$  denotes the matrix inverse to  $L^\alpha_\beta$ . In the case of the coordinate transformation group it is given by  $L^{-1i}_{j'}[\bar{x}] = \delta^i_j \delta(\bar{x}, \bar{x}(\mathbf{x}')) / \partial(\mathbf{x}')$ .

As a result of the action of the group the variables  $\varphi^i$  suffer a transformation which may be expressed in the form

$$\varphi'^i = \Phi^i[\bar{x}, \varphi], \quad (\text{B10})$$

where the function(al)s  $\Phi^i[\bar{x}, \varphi]$  satisfy the identities

$$\Phi^i[e, \varphi] = \varphi^i, \quad (\text{B11})$$

$$\Phi^i[\bar{y}\bar{x}, \varphi] = \Phi^i[\bar{y}, \Phi[\bar{x}, \varphi]]. \quad (\text{B12})$$

Differentiation of (B12) leads to

$$\Phi^i_{,\alpha}[\bar{x}, \varphi] = R^i_\beta[\Phi[\bar{x}, \varphi]] L^{-1\beta}_\alpha[\bar{x}], \quad (\text{B13})$$

where

$$R^i_\alpha[\varphi] \equiv \Phi^i_{,\alpha}[e, \varphi]. \quad (\text{B14})$$

The function(al)s  $R^i_\alpha$  appear in the law of transformation of the  $\varphi$ 's under infinitesimal group operations. Under the action of a group element having the coordinates  $e^\alpha + \delta\xi^\alpha$ , where the  $\delta\xi$ 's are infinitesimal, the  $\varphi$ 's suffer the change

$$\delta\varphi^i = R^i_\alpha \delta\xi^\alpha. \quad (\text{B15})$$

(Functional arguments are again suppressed.) For example, under the infinitesimal coordinate transformation  $\bar{x}^i = x^i + \delta\xi^i$ , the 3-metric  $\gamma_{ij}$  suffers the change

$$\delta\gamma_{ij} = \int R_{ijk'} \delta\xi^{k'} d^3x', \quad (\text{B16})$$

where

$$R_{ijk'} \equiv -\gamma_{ij,k} \delta(\mathbf{x}, \mathbf{x}') - \gamma_{kj,i} \delta(\mathbf{x}, \mathbf{x}') - \gamma_{ik,j} \delta(\mathbf{x}, \mathbf{x}') \quad (\text{B17a})$$

$$= -\delta_{ik',j} - \delta_{jk',i}, \quad (\text{B17b})$$

$$\delta_{ij'} \equiv \gamma_{ik} \delta^{k'}. \quad (\text{B18})$$

If the transformation laws (B10) and (B14) are linear as in this case, then  $R^i_{\alpha, jk} = 0$ . (The reader should avoid confusing differentiation with respect to the  $x$ 's in the explicit notation and functional differentiation with respect to the  $\varphi$ 's in the compact notation.)

With these preliminaries out of the way the question of imposing a metric on the manifold of  $\varphi$ 's may now be considered. Let such a metric be denoted by  $g_{ij}[\varphi]$ . If the group is to generate isometric motions in the manifold then this metric must satisfy Killing's equation:

$$\delta g_{ij,k} \delta\varphi^k + g_{kj} \delta\varphi^k, i + g_{ik} \delta\varphi^k, j = 0, \quad (\text{B19})$$

with  $\delta\varphi^i$  given by Eq. (B15). It is not difficult to see that this equation may be regarded as a group transformation law for  $g_{ij}$ :

$$\delta g_{ij} \equiv g_{ij,k} \delta\varphi^k = -g_{kj} R^k_{\alpha,i} \delta\xi^\alpha - g_{ik} R^k_{\alpha,j} \delta\xi^\alpha. \quad (\text{B20})$$

When the transformation law (B15) is linear and homogeneous, then

$$\delta\varphi^i = R^i_{\alpha,j} \varphi^j \delta\xi^\alpha, \quad (\text{B21})$$

and Eq. (B20) says simply that  $g_{ij}$  must transform contragrediently to the Kronecker product  $\varphi^i \varphi^j$ . This is a necessary and sufficient condition for the isometry of group operations.

Now let  $d\varphi^i$  be an arbitrary displacement, with the corresponding "arc length"  $ds$  given by

$$ds^2 = g_{ij} d\varphi^i d\varphi^j. \quad (\text{B22})$$

If  $d\varphi^i$  happens to be orthogonal to the orbit of  $\varphi$  under the group then it satisfies the condition

$$R^j_{\alpha} g_{ij} d\varphi^i = 0, \quad (\text{B23})$$

and (B22) gives directly the distance between neighboring orbits. In the case of the manifold  $M^{\infty 3}$ , with the metric (6.5) and the transformation law (B16), (B17), the condition (B23) takes the form

$$\gamma^{jk} (d\gamma_{ij,k} - d\gamma_{jk,i}) = 0. \quad (\text{B24})$$

More generally, the distance between  $\text{orb}\varphi$  and  $\text{orb}(\varphi + d\varphi)$  is given by

$$d\tilde{s}^2 = g_{ij} \tilde{d}\varphi^i \tilde{d}\varphi^j, \quad (\text{B25})$$

where  $\tilde{d}\varphi^i$  is the projection of  $d\varphi^i$  normal to the orbit:

$$\tilde{d}\varphi^i = (\delta^i_j - R^i_\alpha \gamma^{\alpha\beta} R^k_\beta g_{kj}) d\varphi^j, \quad (\text{B26})$$

$$\gamma_{\alpha\gamma} \gamma^{\gamma\beta} = \delta_{\alpha}^{\beta}, \quad (\text{B27})$$

$$\gamma_{\alpha\beta} \equiv g_{ij} R^i_\alpha R^j_\beta. \quad (\text{B28})$$

When the indices  $\alpha, \beta$  include continuous labels the "matrix"  $\gamma_{\alpha\beta}$  is typically a differential operator (sum of differentiated  $\delta$  functions) and its inverse  $\gamma^{\alpha\beta}$  is a Green's function.

Equation (B25) may be written in the alternative forms

$$d\tilde{s}^2 = \tilde{g}_{ij} d\varphi^i d\varphi^j = \tilde{g}_{ij} \tilde{d}\varphi^i \tilde{d}\varphi^j, \quad (\text{B29})$$

where

$$\tilde{g}_{ij} \equiv g_{ij} - g_{ik} R^k_\alpha \gamma^{\alpha\beta} R^l_\beta g_{lj}. \quad (\text{B30})$$

$\tilde{g}_{ij}$  is the metric in the manifold of orbits, and the question arises how it transforms under the group. This question is answered by establishing the following

transformation laws:

$$\delta R^i_\alpha \equiv R^i_{\alpha,j} \delta \varphi^j = (R^i_{\beta,j} R^i_\alpha - c^\gamma_{\beta\alpha} R^i_\gamma) \delta \xi^\beta, \quad (B31)$$

$$\delta \gamma_{\alpha\beta} \equiv \gamma_{\alpha\beta, i} \delta \varphi^i = - (c^\gamma_{\delta\alpha} \gamma_{\gamma\beta} + c^\gamma_{\delta\beta} \gamma_{\alpha\gamma}) \delta \xi^\delta, \quad (B32)$$

$$\delta \gamma^{\alpha\beta} \equiv \gamma^{\alpha\beta, i} \delta \varphi^i = (c^\alpha_{\delta\gamma} \gamma^{\gamma\beta} + c^\beta_{\delta\gamma} \gamma^{\alpha\gamma}) \delta \xi^\delta. \quad (B33)$$

Equations (B32) and (B33) are corollaries of (B19), (B27), and (B31), while (B31) itself is a consequence of the identity

$$R^i_{\alpha,j} R^j_\beta - R^i_{\beta,j} R^j_\alpha = R^i_\gamma c^\gamma_{\alpha\beta}, \quad (B34)$$

which is obtained by differentiating Eq. (B13) with respect to  $\tilde{x}^\beta$ , setting  $\tilde{x} = e$ , antisymmetrizing in  $\alpha$  and  $\beta$ , and making use of (B8). With the aid of (B31) and (B33) it is straightforward to show that  $\bar{g}_{ij}$  transforms just like  $g_{ij}$ . This means that group operations are isometries of  $\bar{g}_{ij}$  just as they are of  $g_{ij}$ , and suggests that  $\bar{g}_{ij}$  is effectively a function of  $\text{orb}\varphi$  alone. In order to make this fact explicit a coordinatization of the orbit manifold will be introduced.

This may be accomplished by first introducing a hypersurface in the  $\varphi$  manifold, defined by a set of simultaneous equations

$$f_\alpha[\varphi] = 0, \quad (B35)$$

where the index  $\alpha$  ranges over the same continuum (or discrete set, as the case may be) as the group indices. The only requirement on the hypersurface is that it intersect the orbit of every point contained in (at least) some finite portion of the  $\varphi$  manifold. A coordinate system is then laid down in this hypersurface, with the coordinates denoted by  $z^A$ . If the hypersurface has been carefully chosen each orbit will intersect it in a single point, and the  $z$ 's at that point may be used to label the orbit itself. For example, in the manifold  $M^{3^3}$  one may choose for the equations (B35) the harmonic condition  $(\gamma^{1/2} \gamma^{ij})_{,j} = 0$ ; then any three of the functions  $\varphi^{AB}(\eta)$  of Eq. (5.3) may be chosen as the  $z$ 's.

A general point in the  $\varphi$  manifold will be reached by moving off the hypersurface along (i.e., within) an orbit thus:

$$\varphi^i[\tilde{x}, z] = \Phi^i[\tilde{x}, \varphi_0[z]], \quad (B36)$$

where  $\varphi_0[z]$  is the starting point on the hypersurface. The group coordinates  $\tilde{x}^\alpha$  together with the  $z$ 's provide a new labeling scheme for the points of the  $\varphi$  manifold, and the task before us is to compute the metrics  $g_{ij}$  and  $\bar{g}_{ij}$  in this new coordinate system. For this purpose we shall need the relations

$$\varphi^i_{,\alpha} = R^i_{\beta,\alpha}[\varphi] L^{-1\beta} \alpha[\tilde{x}], \quad (B37)$$

$$\varphi^i_{,\alpha A} = R^i_{\beta,i}[\varphi] \varphi^j_{,\alpha} L^{-1\beta} \alpha[\tilde{x}], \quad (B38)$$

$$\varphi^i_{,\alpha\beta} = R^i_{\gamma,i}[\varphi] R^j_\delta[\varphi] L^{-1\gamma} \alpha[\tilde{x}] L^{-1\delta} \beta[\tilde{x}] + R^i_\gamma[\varphi] L^{-1\gamma} \alpha, \quad (B39)$$

which are obtained by applying Eq. (B13) to (B36). In the work which follows the arguments  $\tilde{x}$ ,  $\varphi$ , and  $z$  will be suppressed.

It is straightforward to compute

$$g_{\alpha\beta} \equiv g_{ij} \varphi^i_{,\alpha} \varphi^j_{,\beta} = \gamma_{\gamma\delta} L^{-1\gamma} \alpha L^{-1\delta} \beta, \quad (B40)$$

$$g_{\alpha A} = g_{A\alpha} \equiv g_{ij} \varphi^i_{,\alpha} \varphi^j_{,A} = g_{ij} R^i_\beta L^{-1\beta} \alpha_{,A}, \quad (B41)$$

$$g_{AB} \equiv g_{ij} \varphi^i_{,A} \varphi^j_{,B} = \bar{g}_{AB} + g_{A\alpha} \bar{g}^{\alpha\beta} g_{B\beta}, \quad (B42)$$

where

$$\bar{g}_{AB} \equiv \bar{g}_{ij} \varphi^i_{,A} \varphi^j_{,B}, \quad (B43)$$

$$\bar{g}^{\alpha\beta} \equiv L^\alpha_\gamma L^\beta_\delta \gamma^{\gamma\delta}, \quad g_{\alpha\beta} \bar{g}^{\alpha\beta} = \delta_{\alpha\beta}. \quad (B44)$$

One then readily verifies that the contravariant metric, with components  $g^{\alpha\beta}$ ,  $g^{\alpha A} (= g^{AA})$ ,  $g^{AB}$ , is given by

$$g^{\alpha\beta} = \bar{g}^{\alpha\beta} + \bar{g}^{\alpha\gamma} g_{\gamma A} g^{AB} g_{B\delta} \bar{g}^{\delta\beta}, \quad (B45)$$

$$g^{\alpha A} = -\bar{g}^{\alpha\beta} g_{B\beta} g^{BA}, \quad (B46)$$

$$\bar{g}_{AC} g^{CB} = \delta_{A}{}^B. \quad (B47)$$

It is now easy to show that  $g_{AB}$  and  $\bar{g}_{AB}$  (and hence  $g^{AB}$ ) are independent of the  $x$ 's. Thus, using (B37) and (B38) one finds

$$\begin{aligned} g_{AB,\alpha} &= g_{ij,k} \varphi^i_{,\alpha} \varphi^j_{,B} \varphi^k_{,\alpha} + g_{ij} (\varphi^i_{,\alpha\alpha} \varphi^j_{,B} + \varphi^i_{,\alpha} \varphi^j_{,B\alpha}) \\ &= (g_{ij,k} R^k_\beta + g_{kj} R^k_{\beta,i} + g_{ik} R^k_{\beta,j}) \\ &\quad \times \varphi^i_{,\alpha} \varphi^j_{,B} L^{-1\beta} \alpha. \end{aligned} \quad (B48)$$

But this expression vanishes by virtue of the group transformation law for  $g_{ij}$  [Eq. (B20)]. Since  $\bar{g}_{ij}$  obeys the same transformation law it follows that

$$\bar{g}_{AB,\alpha} = 0. \quad (B49)$$

That is, the metric  $\bar{g}_{AB}$  of the orbit manifold depends only on the  $z$ 's, as was expected.

For the study of geodesics in the  $\varphi$  manifold the following derivatives will also be needed:

$$\begin{aligned} g_{\alpha\beta,\gamma} &= g_{ij,k} \varphi^i_{,\alpha} \varphi^j_{,\beta} \varphi^k_{,\gamma} + g_{ij} (\varphi^i_{,\alpha\gamma} \varphi^j_{,\beta} + \varphi^i_{,\alpha} \varphi^j_{,\beta\gamma}) \\ &= g_{\alpha\delta} L^\delta_\epsilon L^{-1\epsilon} \gamma, \quad (B50) \end{aligned}$$

$$\begin{aligned} g_{\alpha A,\beta} &= g_{ij,k} \varphi^i_{,\alpha} \varphi^j_{,\beta} \varphi^k_{,\alpha} + g_{ij} (\varphi^i_{,\alpha\beta} \varphi^j_{,\alpha} + \varphi^i_{,\alpha} \varphi^j_{,\alpha\beta}) \\ &= g_{A\gamma} L^\gamma_\delta L^{-1\delta} \beta, \quad (B51) \end{aligned}$$

These are obtained with the aid of (B19), (B37), (B38), and (B39).

The geodesic equations in the  $\varphi$  manifold may be written in the form

$$\begin{aligned} 0 = & g_{\alpha\beta} \frac{d^2 \tilde{x}^\beta}{ds^2} + g_{\alpha A} \frac{d^2 z^A}{ds^2} + \Gamma_{\beta\gamma\alpha} \frac{d\tilde{x}^\beta}{ds} \frac{d\tilde{x}^\gamma}{ds} + 2\Gamma_{\beta A\alpha} \frac{d\tilde{x}^\beta}{ds} \frac{dz^A}{ds} \\ & + \Gamma_{AB\alpha} \frac{dz^A}{ds} \frac{dz^B}{ds}, \end{aligned} \quad (B52)$$

$$\begin{aligned} 0 = & g_{A\alpha} \frac{d^2 \tilde{x}^\alpha}{ds^2} + g_{AB} \frac{d^2 z^B}{ds^2} + \Gamma_{\alpha\beta A} \frac{d\tilde{x}^\alpha}{ds} \frac{d\tilde{x}^\beta}{ds} + 2\Gamma_{\alpha BA} \frac{d\tilde{x}^\alpha}{ds} \frac{dz^B}{ds} \\ & + \Gamma_{BCA} \frac{dz^B}{ds} \frac{dz^C}{ds}, \end{aligned} \quad (B53)$$

where the  $\Gamma$ 's are the Christoffel symbols. Multiplying

(B52) by  $g_{A\delta}\bar{g}^{\delta\alpha}$ , subtracting the result from (B53) and using the fact that  $g_{AB,\alpha}=0$ , one obtains

$$\begin{aligned} 0 = & \bar{g}_{AB} \frac{d^2 z^B}{ds^2} + \bar{\Gamma}_{BCA} \frac{dz^B}{ds} \frac{dz^C}{ds} + \frac{1}{2} [g_{\alpha A,\beta} + g_{\beta A,\alpha} - g_{\alpha\beta,A} - g_{A\delta}\bar{g}^{\delta\gamma}(g_{\alpha\gamma,\beta} + g_{\beta\gamma,\alpha} - g_{\alpha\beta,\gamma})] \frac{d\bar{x}^\alpha}{ds} \frac{d\bar{x}^\beta}{ds} \\ & + [g_{\alpha A,B} - g_{\alpha B,A} - g_{A\gamma}\bar{g}^{\gamma\beta}(g_{\alpha\beta,B} + g_{\beta B,\alpha} - g_{\alpha B,\beta})] \frac{d\bar{x}^\alpha}{ds} \frac{dz^B}{ds} + \frac{1}{2} [(g_{B\alpha}\bar{g}^{\alpha\beta}g_{\beta A}),_C + (g_{C\alpha}\bar{g}^{\alpha\beta}g_{\beta A}),_B - (g_{B\alpha}\bar{g}^{\alpha\beta}g_{\beta C}),_A \\ & - g_{A\beta}\bar{g}^{\beta\alpha}(g_{B\alpha,C} + g_{C\alpha,B})] \frac{dz^B}{ds} \frac{dz^C}{ds}, \end{aligned} \quad (\text{B54})$$

where the  $\bar{\Gamma}$ 's are the Christoffel symbols of the orbit manifold. The terms of this equation can be regrouped by judicious use of identities such as  $g_{\alpha B,A} = (g_{\alpha\gamma}\bar{g}^{\gamma\delta}g_{\delta B}),_A$  and  $\bar{g}^{\gamma\delta},_A g_{\delta\beta} = -\bar{g}^{\gamma\delta}g_{\beta\delta},_A$  and by replacing derivatives of the form  $g_{\alpha\beta,\gamma}$  and  $g_{\alpha A,\beta}$  by their expressions (B50), (B51). The final useful result is

$$\begin{aligned} 0 = & \bar{g}_{AB} \frac{d^2 z^B}{ds^2} + \bar{\Gamma}_{BCA} \frac{dz^B}{ds} \frac{dz^C}{ds} + [(g_{A\gamma}\bar{g}^{\gamma\alpha}),_C - (g_{C\gamma}\bar{g}^{\gamma\alpha}),_A] \frac{dz^C}{ds} \left( g_{\alpha\beta} \frac{d\bar{x}^\beta}{ds} + g_{\alpha B} \frac{dz^B}{ds} \right) \\ & + \frac{1}{2} \bar{g}^{\gamma\delta},_A \left( g_{\gamma\alpha} \frac{d\bar{x}^\alpha}{ds} + g_{\gamma B} \frac{dz^B}{ds} \right) \left( g_{\delta\beta} \frac{d\bar{x}^\beta}{ds} + g_{\delta C} \frac{dz^C}{ds} \right) + g_{A\gamma}\bar{g}^{\gamma\beta}L^\delta_\epsilon(L^{-1\epsilon}_{\beta,\alpha} - L^{-1\epsilon}_{\alpha,\beta}) \frac{d\bar{x}^\alpha}{ds} \left( g_{\delta\epsilon} \frac{d\bar{x}^\epsilon}{ds} + g_{\delta B} \frac{dz^B}{ds} \right). \end{aligned} \quad (\text{B55})$$

Now suppose the geodesic intersects one of the orbits orthogonally. The condition for this is [cf. Eq. (B23)]

$$R^i{}_\beta g_{ij} d\varphi^j / ds = 0, \quad (\text{B56})$$

which, when multiplied by  $L^{-1\beta}_\alpha$ , yields

$$g_{\alpha\beta} \frac{d\bar{x}^\beta}{ds} + g_{\alpha A} \frac{dz^A}{ds} = 0. \quad (\text{B57})$$

When this condition is satisfied we have

$$g_{AB} \frac{dz^A}{ds} \frac{dz^B}{ds} + 2g_{\alpha A} \frac{d\bar{x}^\alpha}{ds} \frac{dz^A}{ds} + g_{\alpha\beta} \frac{d\bar{x}^\alpha}{ds} \frac{d\bar{x}^\beta}{ds} + g_{\alpha A} \frac{dz^A}{ds} \frac{dz^B}{ds} = \bar{g}_{AB} \frac{dz^A}{ds} \frac{dz^B}{ds}, \quad (\text{B58})$$

and hence

$$ds^2 = d\bar{s}^2, \quad (\text{B59})$$

so that the arc length in the  $\varphi$  manifold becomes the same as in the orbit manifold. Moreover, by virtue of (B55) it follows that the  $z$ 's in this case satisfy also the geodesic equation in the orbit manifold,

$$\bar{g}_{AB} \frac{d^2 z^B}{d\bar{s}^2} + \bar{\Gamma}_{BCA} \frac{dz^B}{d\bar{s}} \frac{dz^C}{d\bar{s}} = 0, \quad (\text{B60})$$

provided the orthogonality condition (B57) is maintained along the entire length of the geodesic. But this is an immediate consequence of Eqs. (B50), (B51), and (B52), for by differentiating the left-hand side of (B57) with respect to  $s$ , one obtains

$$\begin{aligned} g_{\alpha\beta} \frac{d^2 \bar{x}^\beta}{ds^2} + g_{\alpha A} \frac{d^2 z^A}{ds^2} + g_{\alpha\beta,\gamma} \frac{d\bar{x}^\beta}{ds} \frac{d\bar{x}^\gamma}{ds} + (g_{\alpha\beta,A} + g_{\alpha A,\beta}) \frac{d\bar{x}^\beta}{ds} \frac{dz^A}{ds} + g_{\alpha A,B} \frac{dz^A}{ds} \frac{dz^B}{ds} \\ = \frac{1}{2} g_{\beta\gamma,\alpha} \frac{d\bar{x}^\beta}{ds} \frac{d\bar{x}^\gamma}{ds} + g_{\beta A,\alpha} \frac{d\bar{x}^\beta}{ds} \frac{dz^A}{ds} = \left( g_{\delta\beta} \frac{d\bar{x}^\beta}{ds} + g_{\delta A} \frac{dz^A}{ds} \right) L^\delta_\epsilon L^{-1\epsilon}_{\alpha,\gamma} \frac{d\bar{x}^\gamma}{ds}, \end{aligned} \quad (\text{B61})$$

which vanishes by virtue of (B57) itself. Therefore, if the geodesic intersects one orbit orthogonally then it intersects every orbit in its path orthogonally, and, moreover, it traces out a geodesic curve in the orbit manifold.

## A MODEL OF LEPTONS\*

Steven Weinberg†

Laboratory for Nuclear Science and Physics Department,  
Massachusetts Institute of Technology, Cambridge, Massachusetts  
(Received 17 October 1967)

Leptons interact only with photons, and with the intermediate bosons that presumably mediate weak interactions. What could be more natural than to unite<sup>1</sup> these spin-one bosons into a multiplet of gauge fields? Standing in the way of this synthesis are the obvious differences in the masses of the photon and intermediate meson, and in their couplings. We might hope to understand these differences by imagining that the symmetries relating the weak and electromagnetic interactions are exact symmetries of the Lagrangian but are broken by the vacuum. However, this raises the specter of unwanted massless Goldstone bosons.<sup>2</sup> This note will describe a model in which the symmetry between the electromagnetic and weak interactions is spontaneously broken, but in which the Goldstone bosons are avoided by introducing the photon and the intermediate-boson fields as gauge fields.<sup>3</sup> The model may be renormalizable.

We will restrict our attention to symmetry groups that connect the observed electron-type leptons only with each other, i.e., not with muon-type leptons or other unobserved leptons or hadrons. The symmetries then act on a left-handed doublet

$$L \equiv [\frac{1}{2}(1 + \gamma_5)] \begin{pmatrix} \nu & e \\ e & \bar{e} \end{pmatrix} \quad (1)$$

$$\mathcal{L} = -\frac{1}{4}(\partial_\mu \vec{A}_\nu - \partial_\nu \vec{A}_\mu + g \vec{A}_\mu \times \vec{A}_\nu)^2 - \frac{1}{4}(\partial_\mu B_\nu - \partial_\nu B_\mu)^2 - \bar{R} \gamma^\mu (\partial_\mu - ig' B_\mu) R - L \gamma^\mu (\partial_\mu - ig \vec{\tau} \cdot \vec{A}_\mu - i \frac{1}{2} g' B_\mu) L$$

$$-\frac{1}{2}[\partial_\mu \varphi - ig \vec{A}_\mu \cdot \vec{\tau} \varphi + i \frac{1}{2} g' B_\mu \varphi]^2 - G_e (\bar{L} \varphi R + \bar{R} \varphi^\dagger L) - M_1^2 \varphi^\dagger \varphi + h(\varphi^\dagger \varphi)^2. \quad (4)$$

We have chosen the phase of the  $R$  field to make  $G_e$  real, and can also adjust the phase of the  $L$  and  $Q$  fields to make the vacuum expectation value  $\lambda \equiv \langle \varphi^0 \rangle$  real. The "physical"  $\varphi$  fields are then  $\varphi^-$

and on a right-handed singlet

$$R \equiv [\frac{1}{2}(1 - \gamma_5)]e. \quad (2)$$

The largest group that leaves invariant the kinematic terms  $-\bar{L} \gamma^\mu \partial_\mu L - \bar{R} \gamma^\mu \partial_\mu R$  of the Lagrangian consists of the electronic isospin  $\vec{\tau}$  acting on  $L$ , plus the numbers  $N_L, N_R$  of left- and right-handed electron-type leptons. As far as we know, two of these symmetries are entirely unbroken: the charge  $Q = T_3 - N_R - \frac{1}{2}N_L$ , and the electron number  $N = N_R + N_L$ . But the gauge field corresponding to an unbroken symmetry will have zero mass,<sup>4</sup> and there is no massless particle coupled to  $N$ ,<sup>5</sup> so we must form our gauge group out of the electronic isospin  $\vec{\tau}$  and the electronic hypercharge  $Y \equiv N_R + \frac{1}{2}N_L$ .

Therefore, we shall construct our Lagrangian out of  $L$  and  $R$ , plus gauge fields  $\vec{A}_\mu$  and  $B_\mu$  coupled to  $\vec{\tau}$  and  $Y$ , plus a spin-zero doublet

$$\varphi = \begin{pmatrix} \varphi^0 \\ \varphi^- \end{pmatrix} \quad (3)$$

whose vacuum expectation value will break  $\vec{\tau}$  and  $Y$  and give the electron its mass. The only renormalizable Lagrangian which is invariant under  $\vec{\tau}$  and  $Y$  gauge transformations is

and

$$\varphi_1 \equiv (\varphi^0 + \varphi^{0\dagger} - 2\lambda)/\sqrt{2} \quad \varphi_2 \equiv (\varphi^0 - \varphi^{0\dagger})/i\sqrt{2}. \quad (5)$$

The condition that  $\varphi_1$  have zero vacuum expectation value to all orders of perturbation theory tells us that  $\lambda^2 \cong M_1^2/2h$ , and therefore the field  $\varphi_1$  has mass  $M_1$  while  $\varphi_2$  and  $\varphi^-$  have mass zero. But we can easily see that the Goldstone bosons represented by  $\varphi_2$  and  $\varphi^-$  have no physical coupling. The Lagrangian is gauge invariant, so we can perform a combined isospin and hypercharge gauge transformation which eliminates  $\varphi^-$  and  $\varphi_2$  everywhere<sup>6</sup> without changing anything else. We will see that  $G_e$  is very small, and in any case  $M_1$  might be very large,<sup>7</sup> so the  $\varphi_1$  couplings will also be disregarded in the following.

The effect of all this is just to replace  $\varphi$  everywhere by its vacuum expectation value

$$\langle \varphi \rangle = \lambda \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \quad (6)$$

The first four terms in  $\mathcal{L}$  remain intact, while the rest of the Lagrangian becomes

$$\begin{aligned} & -\frac{1}{8}\lambda^2 g^2 [(A_\mu^{-1})^2 + (A_\mu^{-2})^2] \\ & -\frac{1}{8}\lambda^2 (gA_\mu^3 + g'B_\mu)^2 - \lambda G_e \bar{e}e. \end{aligned} \quad (7)$$

We see immediately that the electron mass is  $\lambda G_e$ . The charged spin-1 field is

$$W_\mu \equiv 2^{-1/2}(A_\mu^{-1} + iA_\mu^{-2}) \quad (8)$$

and has mass

$$M_W = \frac{1}{2}\lambda g. \quad (9)$$

The neutral spin-1 fields of definite mass are

$$Z_\mu = (g^2 + g'^2)^{-1/2}(gA_\mu^3 + g'B_\mu), \quad (10)$$

$$A_\mu = (g^2 + g'^2)^{-1/2}(-g'A_\mu^3 + gB_\mu). \quad (11)$$

Their masses are

$$M_Z = \frac{1}{2}\lambda(g^2 + g'^2)^{1/2}, \quad (12)$$

$$M_A = 0, \quad (13)$$

so  $A_\mu$  is to be identified as the photon field. The interaction between leptons and spin-1 mesons is

$$\begin{aligned} & \frac{ig}{2\sqrt{2}} \bar{e} \gamma^\mu (1 + \gamma_5)^\nu W_\mu + \text{H.c.} + \frac{igg'}{(g^2 + g'^2)^{1/2}} \bar{e} \gamma^\mu e A_\mu \\ & + \frac{i(g^2 + g'^2)^{1/2}}{4} \left[ \left( \frac{3g'^2 - g^2}{g'^2 + g^2} \right) \bar{e} \gamma^\mu e - \bar{e} \gamma^\mu \gamma_5 e + \bar{\nu} \gamma^\mu (1 + \gamma_5)^\nu \right] Z_\mu. \end{aligned} \quad (14)$$

We see that the rationalized electric charge is

$$e = gg'/(g^2 + g'^2)^{1/2} \quad (15)$$

and, assuming that  $W_\mu$  couples as usual to hadrons and muons, the usual coupling constant of weak interactions is given by

$$G_W/\sqrt{2} = g^2/8M_W^2 = 1/2\lambda^2. \quad (16)$$

Note that then the  $e-\varphi$  coupling constant is

$$G_e = M_e/\lambda = 2^{1/4} M_e G_W^{1/2} = 2.07 \times 10^{-6}.$$

The coupling of  $\varphi_1$  to muons is stronger by a factor  $M_\mu/M_e$ , but still very weak. Note also that (14) gives  $g$  and  $g'$  larger than  $e$ , so (16) tells us that  $M_W > 40$  BeV, while (12) gives  $M_Z > M_W$  and  $M_Z > 80$  BeV.

The only unequivocal new predictions made

by this model have to do with the couplings of the neutral intermediate meson  $Z_\mu$ . If  $Z_\mu$  does not couple to hadrons then the best place to look for effects of  $Z_\mu$  is in electron-neutron scattering. Applying a Fierz transformation to the  $W$ -exchange terms, the total effective  $e-\nu$  interaction is

$$\frac{G_W}{\sqrt{2}} \bar{\nu} \gamma_\mu (1 + \gamma_5)^\nu \left\{ \frac{(3g^2 - g'^2)}{2(g^2 + g'^2)} \bar{e} \gamma^\mu e + \frac{3}{2} \bar{e} \gamma^\mu \gamma_5 e \right\}.$$

If  $g \gg e$  then  $g \gg g'$ , and this is just the usual  $e-\nu$  scattering matrix element times an extra factor  $\frac{3}{2}$ . If  $g \approx e$  then  $g \ll g'$ , and the vector interaction is multiplied by a factor  $-\frac{1}{2}$  rather than  $\frac{3}{2}$ . Of course our model has too many arbitrary features for these predictions to be

taken very seriously, but it is worth keeping in mind that the standard calculation<sup>3</sup> of the electron-neutrino cross section may well be wrong.

Is this model renormalizable? We usually do not expect non-Abelian gauge theories to be renormalizable if the vector-meson mass is not zero, but our  $Z_\mu$  and  $W_\mu$  mesons get their mass from the spontaneous breaking of the symmetry, not from a mass term put in at the beginning. Indeed, the model Lagrangian we start from is probably renormalizable, so the question is whether this renormalizability is lost in the reordering of the perturbation theory implied by our redefinition of the fields. And if this model is renormalizable, then what happens when we extend it to include the couplings of  $A_\mu$  and  $B_\mu$  to the hadrons?

I am grateful to the Physics Department of MIT for their hospitality, and to K. A. Johnson for a valuable discussion.

\*This work is supported in part through funds provided by the U. S. Atomic Energy Commission under Contract No. AT(30-1)2098.

†On leave from the University of California, Berkeley, California.

<sup>1</sup>The history of attempts to unify weak and electromagnetic interactions is very long, and will not be reviewed here. Possibly the earliest reference is E. Fer-

mi, Z. Physik 88, 161 (1934). A model similar to ours was discussed by S. Glashow, Nucl. Phys. 22, 579 (1961); the chief difference is that Glashow introduces symmetry-breaking terms into the Lagrangian, and therefore gets less definite predictions.

<sup>2</sup>J. Goldstone, Nuovo Cimento 19, 154 (1961); J. Goldstone, A. Salam, and S. Weinberg, Phys. Rev. 127, 965 (1962).

<sup>3</sup>P. W. Higgs, Phys. Letters 12, 132 (1964), Phys. Rev. Letters 13, 508 (1964), and Phys. Rev. 145, 1156 (1966); F. Englert and R. Brout, Phys. Rev. Letters 13, 321 (1964); G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble, Phys. Rev. Letters 13, 585 (1964).

<sup>4</sup>See particularly T. W. B. Kibble, Phys. Rev. 155, 1554 (1967). A similar phenomenon occurs in the strong interactions; the  $\rho$ -meson mass in zeroth-order perturbation theory is just the bare mass, while the  $A_1$  meson picks up an extra contribution from the spontaneous breaking of chiral symmetry. See S. Weinberg, Phys. Rev. Letters 18, 507 (1967), especially footnote 7; J. Schwinger, Phys. Letters 24B, 473 (1967); S. Glashow, H. Schnitzer, and S. Weinberg, Phys. Rev. Letters 19, 139 (1967), Eq. (13) et seq.

<sup>5</sup>T. D. Lee and C. N. Yang, Phys. Rev. 98, 101 (1955).

<sup>6</sup>This is the same sort of transformation as that which eliminates the nonderivative  $\tilde{\pi}$  couplings in the  $\sigma$  model; see S. Weinberg, Phys. Rev. Letters 18, 188 (1967). The  $\tilde{\pi}$  reappears with derivative coupling because the strong-interaction Lagrangian is not invariant under chiral gauge transformation.

<sup>7</sup>For a similar argument applied to the  $\sigma$  meson, see Weinberg, Ref. 6.

<sup>8</sup>R. P. Feynman and M. Gell-Mann, Phys. Rev. 109, 193 (1957).



CONSTRUCTION OF A CROSSING SYMMETRIC, REGGE BEHAVED AMPLITUDE FOR  
LINEARLY RISING TRAJECTORIES

G. Veneziano <sup>\*+</sup>

CERN - Geneva

A B S T R A C T

A representation of the scattering amplitude, containing an average Regge behaviour and crossing symmetry for linearly rising trajectories, is proposed. It obeys superconvergence sum rules at all  $t$ , exhibits in a clear way the Regge poles vs. resonances duality and demands families of parallel daughters.

---

\* On leave of absence from the Weizmann Institute of Science, Rehovoth, Israel

+ Address after 1 September 1968: Department of Physics, M.I.T., Cambridge, Mass. USA

Crossing has been the first ingredient used to make Regge theory a predictive concept in high energy physics. However, a complete and satisfactory way of imposing crossing and crossed channel unitarity is still lacking. We can look at the recent investigations on the properties of Reggeization at  $t=0$  as giving a first encouraging set of results along this line of thinking <sup>1)</sup>. A technically different approach, based on superconvergence, has been also recently investigated <sup>2)</sup>, and the possibility of a self-consistent determination of the physical parameters, through the use of sum rules, has been stressed.

In this note we propose a quite simple expression for the relativistic scattering amplitude, that obeys the requirements of Regge asymptotics and crossing symmetry in the case of linearly rising trajectories. Its explicit form is suggested by the work of Ref. <sup>3)</sup> and contains only a few free parameters <sup>\*</sup>.

Our expression contains automatically Regge poles in families of parallel trajectories (at all  $t$ ) with residue in definite ratios. It furthermore satisfies the conditions of superconvergence <sup>4)</sup> and exhibits in a nice fashion the duality between Regge poles and resonances in the scattering amplitude.

The first example we want to discuss is the scattering  $\pi\pi \rightarrow \pi\omega$ , whose convenient properties have been already stressed in Ref. <sup>3)</sup>. We introduce the invariant amplitude  $A(s,t,u)$  through the definition of the  $T$  matrix

$$T = \epsilon_{\mu\nu\rho\sigma} \epsilon_{\rho_1} \epsilon_{\nu} \epsilon_{\rho_2} \epsilon_{\sigma} \cdot A(s,t,u) \quad (1)$$

<sup>\*</sup>) We shall mostly work here in the approximation of real, linear trajectories and consequently of narrow resonances. We briefly discuss the effects of a non-zero imaginary part in the trajectory function which, in any case, we demand to have a linearly rising real part.

where  $P_i$  are the pion momenta and  $e_\mu$  is the  $\omega$  polarization vector.  $A(s,t,u)$  has only dynamical singularities as it is free of kinematical ones. It is also completely symmetric in the three Mandelstam variables.

It was found in Ref. 3) that a "good" parametrization of  $A$  at high  $s$  and fixed  $t$  could be written as:

$$A(s,t,u) \underset{s \rightarrow \infty}{\simeq} \frac{\bar{\beta}}{\pi} \Gamma(1-\alpha(t)) (-\alpha(s))^{\alpha(t)-1} + (s \leftrightarrow u) \quad (2)$$

with  $\bar{\beta}$  = constant. We use the word "good" in the sense that Eq. (2), when used as an input, is able to reproduce itself quite consistently through the use of superconvergence sum rules.

What is the amplitude for non-asymptotic values of  $s$ ? If Eq. (2) was exact after some  $\bar{s}$ , analyticity in the  $s$  plane (at fixed  $t$ ) would require it to be valid at all  $s$  and Eq. (2) is certainly a solution of superconvergence. However, Eq. (2) does not satisfy  $s,t$  crossing as this demands poles in  $s$  such as those induced in  $t$  by the  $\Gamma(1-\alpha(t))$  factor. On the other hand these poles in  $s$  could in principle destroy the asymptotic behaviour (2) through the introduction of fixed singularities. The lowest moment sum rules are just imposing that this is not happening at the nearest negative integers. Furthermore, we expect that the presence of bumps in the low energy region will produce (through analyticity) a modification of the high energy form which will not be as smooth as Eq. (2), but will rather show oscillations in  $s$ .

Consequently, we take out the factor  $(-\alpha(s))^{\alpha(t)-1}$  and we symmetrize Eq. (2) multiplying by a factor  $\Gamma(1-\alpha(s))$  and dividing by  $\Gamma(2-\alpha(s)-\alpha(t))$  in order to have the correct asymptotic behaviour. After symmetrization in  $s,t,u$  we have

$$A(s,t,u) = \frac{\bar{\beta}}{\pi} \left[ B(1-\alpha(t), 1-\alpha(s)) + B(1-\alpha(t), 1-\alpha(u)) + B(1-\alpha(s), 1-\alpha(u)) \right] \quad (3)$$

where we have introduced the Euler B function

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$$

Notice that, in Eq. (3),  $\bar{\beta}$  must be a constant if we want to have a Regge-like behaviour which, together with crossing, also demands the  $1/\Gamma(\alpha)$  t dependence of the reduced residue function. Equation (3) in fact is hard to modify if one demands an  $s^{\alpha-1}$  behaviour in all channels. The only simple generalization of Eq. (3) seems to consist in the addition of non-leading and similarly structured terms like  $B(m-\alpha(t), n-\alpha(s))$  with  $m, n \geq 1$ .

We now discuss some properties of Eq. (3) in detail

### 1) Behaviour for large positive s and fixed t

The first two terms (we shall come to the last one in a moment) give:

$$A = \frac{\beta(t)}{\sin \pi \alpha(t)} \left[ \frac{\sin \pi(\alpha(s)+\alpha(t))}{\sin \pi \alpha(s)} \frac{\Gamma(\alpha(s)+\alpha(t)-1)}{\Gamma(\alpha(s))} + \frac{\Gamma(1-\alpha(s))}{\Gamma(2-\alpha(s)-\alpha(t))} \right] \quad (4)$$

The second term is purely real (for positive s) and goes like  $(\alpha(s))^{\alpha(t)-1}$ . The first term is the one that corresponds to the Regge term  $(-\alpha(s))^{\alpha(t)-1}$  and has both a real and an imaginary part. Some trivial algebra shows that, from the whole Eq. (4), we have a real piece like

$$\beta(t) \frac{1 - \cos \pi \alpha(t)}{\sin \pi \alpha(t)} [\alpha(s)]^{\alpha(t)-1}; \quad \beta(t) = \bar{\beta}/\Gamma(\alpha(t)),$$

as in the Regge theory, while the s discontinuity is all contained in the form

$$A \underset{s \rightarrow \infty}{\sim} -\beta(t) \cot \alpha(s) [\alpha(s)]^{\alpha(t)-1} \quad (5)$$

If  $\text{Im } \alpha$  is strictly zero, Eq. (5) gives just poles in  $s$  and  $\text{Im } A$  is a sequence of  $\delta$  functions. If  $\text{Im } \alpha$  is different from zero and, for instance, increases with  $s$  (this happens if the total width does not vary strongly with  $s$ ),  $\text{Im } A$  will describe bumps for moderate values of  $s$ , but will finally tend to  $\beta(t)(\alpha(s))^{\alpha(t)-1}$  as in the Regge theory [this is due to  $\cot \alpha(s) \rightarrow -i$ ]. Of course, the parametrization of Eq. (3) can be taken as such only for linearly rising trajectories in which case  $(\alpha(s))^{\alpha(t)}$  is equivalent to  $(\frac{s}{s_0})^{\alpha(t)}$ . However, we only need a leading term in  $\alpha(s)$  going linearly in  $s$ , and this does not imply  $\text{Im } \alpha = 0$ . If  $\text{Im } \alpha \neq 0$  one probably gets, besides moving poles, other singularities (cuts?) as well.

## 2) Singularities in the various channels

Equation (3) has quite nice analytic features. It has cuts in all the three Mandelstam variables starting from the  $2\pi$  threshold, where  $\alpha$  begins to show an imaginary part. However, if we restrict to real linear trajectories, our expression has only poles whenever  $\alpha$  passes through an integer bigger than 0. Furthermore, because of the  $\Gamma(2-\alpha(s)-\alpha(t))$  denominator, no double pole appears, in the sense that the residue in a pole is a polynomial in the other variable.

At first glance our expression shows poles at even values of  $\alpha$  as well, in contrast with invariance principles. As these are always non-leading terms, one can in general eliminate them by the addition of non-leading expressions as explained in the beginning. More amusing to notice is the fact that, at least in this reaction, the elimination of spurious singularities can be achieved with a single condition on the trajectory  $\alpha(t)$ . Take in fact  $\alpha(t) = 2$ . The residue at the pole, produced there by  $\Gamma(1-\alpha(t))$  is simply proportional to  $\alpha(u) + \alpha(s)$ .

We then demand

$$\alpha(s) + \alpha(u) = 0 \quad \text{for } \alpha(t) = 2 \quad (6)$$

which after some easy algebra gives (always for linear trajectories)

$$\alpha(s) + \alpha(t) + \alpha(u) = 2 \quad (7)$$

Equation (7) can easily be transformed into the prediction

$$\alpha(-2m_p^2 + m_\omega^2 + 3m_\pi^2) = \alpha(-0.53\beta\omega^2) = 0 \quad (8)$$

which was derived in Ref.<sup>2)</sup> from the sum rules. The reader can verify that Eq. (7) is enough to cancel all the undesired poles at the even integer values of  $\alpha$ . A further interesting consequence of Eq. (7) concerns the third term of Eq. (3) which could in principle violate the Regge behaviour. Instead, using (7), that term can be rewritten as:

$$\frac{\beta(t)}{\sin\pi\alpha(s)} \frac{\Gamma(\alpha(s) + \alpha(t) - 1)}{\Gamma(\alpha(s))} \quad (9)$$

which is still Regge behaved. The whole Eq. (3) can be rewritten in the form

$$A = \beta(t) \frac{\Gamma(\alpha(s) + \alpha(t) - 1)}{\Gamma(\alpha(s))} \left[ \frac{1 - e^{i\pi\alpha(s)}}{\sin\pi\alpha(s)} + \frac{1 - e^{-i\pi\alpha(t)}}{\sin\pi\alpha(t)} \right] \quad (10)$$

which shows the automatic cancellation of the poles at the even integer values of  $\alpha$ . By use of (7) one can also write (3) in the very symmetric form

$$A = \frac{\bar{\beta}}{\pi^2} \Gamma(1 - \alpha(s)) \Gamma(1 - \alpha(t)) \Gamma(1 - \alpha(u)) \left[ \sin\pi\alpha(s) + \sin\pi\alpha(t) + \sin\pi\alpha(u) \right] \quad (11)$$

As a second example let us consider the process  $\pi\gamma \rightarrow \pi\rho$ . According to our prescription the invariant amplitude A defined as in Eq. (1) will be given by:

$$A_{\pi\gamma \rightarrow \pi\rho} = \frac{\beta_1}{\pi} \left[ B(1-\alpha'_{A_2}(s), 1-\alpha'_\rho(t)) + B(1-\alpha'_{A_2}(u), 1-\alpha'_\rho(t)) - B(1-\alpha'_{A_2}(s), 1-\alpha'_\rho(u)) \right] \quad (21)$$

where  $\pi\pi \rightarrow \eta\rho$  is the t channel. Imposing to find no poles at even integer value for  $\alpha'_\rho(t)$  we obtain:

$$\alpha'_{A_2}(s) + \alpha'_{A_2}(u) + \alpha'_\rho(t) = 2 \quad (22)$$

Imposing absence of poles at the odd integers for  $\alpha'_{A_2}$  we find again Eq. (22). This demands

$$\alpha'_\rho = \alpha'_{A_2} \quad (23)$$

Using  $m_\rho^2 = 0.6 \text{ GeV}^2$  and Eq. (7) we obtain

$$\alpha'_{A_2}(0) = 1 - \frac{1}{2} \frac{3m_\rho^2 - m_\omega^2 - m_\pi^2 + m_\eta^2}{3m_\rho^2 - m_\omega^2 - 3m_\pi^2} \simeq 0.36 \quad (24)$$

We thus predict  $m_{A_2} = 1350 \text{ MeV}$ .

As a third example one could try to build up a scattering amplitude for  $s+s \rightarrow s+s$  ( $s$  being a scalar particle with the vacuum quantum numbers) and try to ask dominance of a leading trajectory passing by the particle itself. This is seen to be impossible with a positive slope of  $\alpha'$ , since the equation similar to (7) gives  $\alpha'(0) = 1$ .

It is possible to extend the above considerations to the more interesting case of  $\pi\pi$  scattering and to obtain a crossing symmetric amplitude in the approximation of  $f^*$  and  $f$  trajectory dominance and disregarding the Pomeranchuk, according to a now accepted philosophy<sup>8),11)</sup>. We find consistency only if  $\alpha'_f = \alpha'_f = \alpha$  and  $\alpha'(0) \approx 1/3$ . Furthermore, we can predict  $\pi\pi$  scattering lengths in terms of  $g_{\rho\pi\pi}^2$  and obtain (apart from the Pomeranchuk contribution)

$$\alpha_1 = 5/2 \quad \alpha_2 = -1.25 \quad M_\pi^{-1}$$

Further details as well as applications of this scheme to more complicated cases will be considered elsewhere.

The author wishes to acknowledge interesting discussions with D. Amati, S. Fubini and M. Toller.

FOOTNOTES AND REFERENCES

- 1) For a general review of these problems see L. Bertocchi, Proceedings of the Heidelberg international conference on elementary particles, North Holland Pub. Co., (1967).
- 2) Such an approach was proposed independently by M. Ademollo, H.R. Rubinstein, G. Veneziano and M.A. Virasoro, Phys.Rev.Letters 19, 1402 (1967) and Phys.Letters 27B, 99 (1968), and by S. Mandelstam, Phys.Rev. 166, 1539 (1968). Further developments and a number of references to related works can be found in Ref. <sup>3)</sup>.
- 3) M. Ademollo, H.R. Rubinstein, G. Veneziano and M.A. Virasoro, Weizmann Institute preprint (1968), submitted to Phys.Rev.
- 4) For superconvergence we mean both the original sum rules proposed by V. de Alfaro, S. Fubini, G. Furlan and C. Rossetti, Phys.Letters 21, 576 (1966), and the more recent generalized superconvergence (finite energy) sum rules [see Ref. <sup>3)</sup> for detailed references]. A unified treatment of all superconvergence sum rules has been given by S. Fubini, Nuovo Cimento 52A, 224 (1967).
- 5) H.R. Rubinstein, A. Schwimmer, G. Veneziano and M.A. Virasoro, Weizmann Institute preprint (1968), submitted to Phys.Rev.Letters; see also Ref. <sup>3)</sup>.
- 6) G. Cosenza, A. Sciarrino and M. Toller, University of Rome preprint Nr. 158 and Trieste preprint.
- 7) C. Schmid, Phys.Rev.Letters 20, 689 (1968).
- 8) H. Harari, Phys.Rev.Letters 20, 1395 (1968).
- 9) G. Cocconi et al., Phys.Rev. 138B, 165 (1965). Having linearly rising trajectories we are also consistent with the Cerulus-Martin bound. See C.B. Chiu and C.I. Tan, Phys.Rev. 162, 1701 (1967).

- 10) We know that, at  $t = 0$ , a simple (irreducible) Lorentz pole does obey factorization [see Ref. <sup>1)</sup>]. It seems also plausible to conjecture (M. Toller, private communication) that this is the only case in which factorization is fulfilled. Since our expression does not probably correspond to a single Lorentz pole, non-leading terms might be needed in order to have factorization. We thank M. Toller for a discussion on this point.
- 11) It may be, however, that one runs into difficulties in adding the Pomeranchuk contribution at the end in a crossing symmetric way. An alternative interesting possibility would be to consider it as originated somehow by the other trajectories (through their non-resonating parts) and not as an independent object. This problem which certainly requires further study, is closely connected to that of the nature of the Pomeranchuk singularity.

# On the Interpretation of Measurement in Quantum Theory

H. D. Zeh

Institut für Theoretische Physik, Universität Heidelberg, Heidelberg, Germany

Received September 19, 1969

---

*It is demonstrated that neither the arguments leading to inconsistencies in the description of quantum-mechanical measurement nor those “explaining” the process of measurement by means of thermodynamical statistics are valid. Instead, it is argued that the probability interpretation is compatible with an objective interpretation of the wave function.*

---

## 1. INTRODUCTION

The problem of measurement in quantum theory and the related problem of how to describe classical phenomena in the framework of quantum theory have received increased attention during recent years. The various contributions express very different viewpoints, and may roughly be classified as follows:

1. Those emphasizing contradictions obtained when the process of measurement is itself described in terms of quantum theory.<sup>(1)</sup>
2. Those claiming that measurement may well be explained by quantum theory in the sense that “quantum-mechanical noncausality” can be derived from statistical uncertainties inherent in the necessarily macroscopic apparatus of measurement.<sup>(2)</sup>
3. Those introducing new physical concepts like hidden variables.<sup>(3)</sup>

Suggestions of the third group are usually based on the first viewpoint, and are meaningful only if they lead to experimental consequences. These have not been confirmed so far.

A measurement in quantum theory is axiomatically described by means of a Hermitian operator. If the eigenstates of this operator are  $\varphi_n$ , and the state of the measured system  $S$  is  $\varphi = \sum c_n \varphi_n$ , then, according to the axiom, the result of the measurement will, with probability  $|c_n|^2$ , be the corresponding eigenvalue  $a_n$  represented physically by a “pointer position,” i.e., by an appropriate state of the measuring device  $M$ . For the most frequent class of measurements, it is furthermore predicted that any following measurement can be described by assuming  $S$  to be in the state  $\varphi_n$  after the measurement.

When describing the process of measurement as a whole in the framework of quantum theory, it is assumed that the apparatus  $M$  can be described by a wave function  $\phi_\alpha$ , the state of the total system  $M + S$  obeying the Schrödinger equation,

$$\psi(t) = e^{iHt} \phi_\alpha \sum_n c_n \varphi_n = \sum_{n,m,\beta} c_n U_{\alpha\beta}^{nm}(t) \phi_\beta \varphi_m \quad (1)$$

with  $U_{\alpha\beta}^{nm}(0) = \delta_{nm} \delta_{\alpha\beta}$ . As the state of a macroscopic apparatus can be determined only incompletely, there must be a large set of states  $\{\phi\}_0$  compatible with the knowledge about  $M$ . If this set of states is assumed to be independent of the state of  $S$  before measurement, a condition on the coefficients  $U_{\alpha\beta}^{nm}(t)$  can be derived from the requirement that the axiom of measurement be fulfilled in the case  $c_n = \delta_{nm_0}$ , i.e.,  $\varphi = \varphi_{n_0}$ . The interaction must be of the von Neumann type<sup>(4)</sup>

$$U_{\alpha\beta}^{nm}(t) = \delta_{nm} u_{\alpha\beta}^n(t) \quad (2)$$

for all but a negligible measure of states of the set  $\{\phi\}_0$ , and for times  $t$  larger than the duration of the measurement. Furthermore, practically all states  $\sum_\beta u_{\alpha\beta}^n(t) \phi_\beta$  must be members of a set  $\{\phi\}_n$  corresponding to a “pointer position  $n$ ” of  $M$ .

In the case of a general state  $\varphi$ , the final total state now takes the form

$$\psi(t) = \sum_{n,\beta} c_n u_{\alpha\beta}^n(t) \phi_\beta \varphi_n \quad (3)$$

It represents a superposition of different pointer positions. This result is said to be in contradiction to the axiom of measurement, because the latter states that the result of the measurement is one of the states  $\sum_\beta u_{\alpha\beta}^n(t) \phi_\beta \varphi_n$ . It is of course very unsatisfactory to assume that the laws of nature change according to whether or not a physical process is a measurement.

The difficulties arising when a macroscopic system is described by quantum theory can be seen more directly by applying the main axiom of quantum theory, i.e., the superposition principle. If there are two possible pointer positions  $\{\phi\}_{n_1}$  and  $\{\phi\}_{n_2}$ , any superposition  $c_1 \phi_{n_1} + c_2 \phi_{n_2}$  must be a possible state. As such superpositions have never been observed (see Wigner<sup>(5)</sup>) one should at least find dynamical causes for their nonoccurrence. Although recent work<sup>(5)</sup> has shown that dynamical stability conditions in the original sense of Schrödinger's<sup>(6)</sup> have a much wider field of applicability than previously expected, the process of measurement does not, because of the above arguments, belong to this class of phenomena.

## 2. CRITICISM OF STATISTICAL INTERPRETATIONS

Results apparently in contradiction to those of the preceding section have been derived in a series of papers<sup>(2)</sup> which try to make use of the uncertainties in the microscopic properties of the apparatus of measurement. The mathematical concept used in these theories is the density matrix formalism.

A simple example may illustrate such theories. If the density matrix describing  $M$  is  $\sum_{\alpha} p_{\alpha} \phi_{\alpha} \phi_{\alpha}^*$ , the total system is described by

$$\rho(t) = \sum_{\alpha, n, n'} p_{\alpha} c_n c_n^* \phi_{\alpha} \phi_{\alpha}^* \varphi_n \varphi_n^*. \quad (4)$$

For a von Neumann interaction, one obtains

$$\rho(t) = e^{iHt} \rho(0) e^{-iHt} = \sum_{\alpha n n' \beta \beta'} p_{\alpha} c_n c_n^* u_{\alpha \beta}^n(t) u_{\alpha \beta'}^{n'}(t) \phi_{\beta} \phi_{\beta'}^* \varphi_n \varphi_n^*. \quad (5)$$

Provided the coefficients  $u_{\alpha \beta}^n(t)$  possess arbitrarily distributed phases guaranteeing that

$$\sum_{\alpha} p_{\alpha} u_{\alpha \beta}^n(t) u_{\alpha \beta'}^{n'}(t) \approx \delta_{nn'} \delta_{\beta \beta'} \quad (6)$$

(the diagonality in  $\beta \beta'$  is not needed),  $\rho(t)$  becomes

$$\rho(t) \approx \sum_n |c_n|^2 \varphi_n \varphi_n^* \sum_{\beta \beta'} q_{\beta \beta'}^n(t) \phi_{\beta} \phi_{\beta'}^*. \quad (7)$$

This density matrix describes exactly the situation postulated by the axiom of measurement.<sup>(4)</sup>

It is tempting to interpret this result by saying that the statistical uncertainty inherent in the macroscopic apparatus is transferred by means of the interaction to the system  $S$ . This means that the outcome of a measurement, i.e., the pointer position, should be exactly predictable if we knew the microscopic state of  $M$ . Equation (3) demonstrates that this interpretation is wrong.<sup>1</sup>

The contradiction between Eqs. (3) and (7) is—aside from the dubious nature of the statistical assumption—due to a circular argument. The density matrix formalism is itself based upon the axiom of measurement. In order to see this, consider the case of a set of states  $\psi^{(i)} = \sum_n c_n^{(i)} \psi_n$  prepared with probabilities  $p^{(i)}$ . The probability of finding the eigenvalue  $a_n$  is then

$$w_n = \sum_i p^{(i)} |c_n^{(i)}|^2 = \text{tr}\{P_w \rho\} \quad (8)$$

<sup>1</sup> The above example is not identical with any of the theories of Ref. 2. It does not, however, use any additional assumptions. As it leads to a contradiction, one of the assumptions used must be wrong. Some of these theories do not start with an ensemble for the initial state of the apparatus, but assume instead that the “pointer position” is represented by some time average. The latter is then transformed into an ensemble average by means of the ergodic theorem. Interpreted rigorously, these theories would prove that the pointer position fluctuates in time.

if  $P_n = |\psi_n\rangle\langle\psi_n|$ , and  $\rho = \sum_i p^{(i)}\psi^{(i)}\psi^{(i)*}$ . The states  $\psi^{(i)}$  will in general not be linearly independent, although  $\rho$  may of course be expanded quadratically in terms of a complete orthogonal set. The reason for the usefulness of  $\rho$  is that, according to the axiom of measurement, all observable quantities can be expressed as linear-antilinear functionals of the wave function.

For example, the statistical ensemble consisting of equal probabilities of neutrons with spin up and spin down in the  $x$  direction cannot be distinguished by measurement from the analogous ensemble having the spins parallel or antiparallel to the  $y$  direction. Both ensembles, however, can be easily prepared by appropriate versions of the Stern-Gerlach experiment. One is justified in describing both ensembles by the same density matrix as long as the axiom of measurement is accepted. However, the density matrix formalism cannot be a complete description of the ensemble, as the ensemble cannot be rederived from the density matrix. The discrepancy between Eqs. (3) and (7) arises since, on the one hand, Eq. (3) must hold for all but a negligible number of members of the ensemble, whereas Eq. (7) is interpreted as describing an ensemble of states  $\varphi_n \sum_\beta u_\beta^n(t) \phi_\beta$ , i.e., each state being essentially different from (3). Only if the measurement axiom is accepted can these ensembles not be distinguished by subsequent observations.

The circularity is more obvious in some versions which avoid the density matrix formalism (e.g., Rosenfeld<sup>(2)</sup> who made repeated use of the probability interpretation although the latter is to be derived). In such cases, the circular argument is considered a "proof of consistency." This viewpoint cannot be accepted, as it would mean that the secondary observation of the pointer position (by a conscious observer or a second apparatus) is a measurement in the axiomatic sense. It corresponds to the interpretation of measurement due to Heisenberg and von Neumann<sup>(4)</sup> (claiming the arbitrariness of the position of the "*Heisenbergscher Schnitt*"), and does not require any contribution from thermodynamics. Bohm's analysis of the process of measurement,<sup>(7)</sup> however, shows the importance of the amplification of the result of a measurement up to the macroscopic scale, thus leading to a natural position of the "*Heisenbergscher Schnitt*." (Relative phases between microscopically realized pointer positions could still be measured.)

The secondary (macroscopic) observation is significantly different from the primary (microscopic) one, for the physical situation between these two observations is described by the reduced wave function. The macroscopic observation can thus be performed in a reversible way, in contrast to the microscopic one, which must result in the reduction. It is implicitly assumed in applying the density matrix formalism that the macroscopic measurement is accompanied by a reduction of the wave function.

### 3. CONSEQUENCES OF A UNIVERSALLY VALID QUANTUM THEORY

The arguments presented so far were based on the assumption that a macroscopic system (the apparatus of measurement) can be described by a wave function  $\phi$ . It appears that this assumption is not valid, for dynamical reasons:

If two systems are described in terms of basic states  $\phi_{k_1}^{(1)}$  and  $\phi_{k_2}^{(2)}$ , the wave

function of the total system can be written as  $\phi = \sum_{k_1 k_2} c_{k_1 k_2} \phi^{(1)}_{k_1} \phi^{(2)}_{k_2}$ . The case where the subsystems are in definite states ( $\phi = \phi^{(1)} \phi^{(2)}$ ) is therefore an exception. Any sufficiently effective interaction will induce correlations. The effectiveness may be measured by the ratios of the interaction matrix elements and the separation of the corresponding unperturbed energy levels. Macroscopic systems possess extremely dense energy spectra. The level distances, for example, of a rotator with moment of inertia  $1 \text{ g cm}^2$  are of the order  $10^{-42} \text{ eV}$ , which value may be compared with the interaction between two electric dipoles of  $1 \text{ e} \times \text{cm}$  at distance  $R$ ,  $e^2 \times \text{cm}^2/R^3 \approx 10^{-7}(\text{cm}/R)^3 \text{ eV}$ . It must be concluded that macroscopic systems are always strongly correlated in their microscopic states. They still do have uncorrelated macroscopic properties, however, if the summations over  $k_1$  and  $k_2$  are each essentially limited to macroscopically equivalent states.<sup>(8)</sup> Since the interactions between macroscopic systems are effective even at astronomical distances, the only "closed system" is the universe as a whole. The assumption of a closed system  $M + S$  is hence unrealistic on a microscopic scale.

The arguments leading to Eq. (3) can be accepted only if the states  $\phi_x$  are interpreted as those of the "remainder of the universe" including the apparatus of measurement, instead of those of the latter alone. It is of course very questionable to describe the universe by a wave function that obeys a Schrödinger equation. Otherwise, however, there is no inconsistency in measurement, as there is no theory. This assumption is referred to as that of "universal validity of quantum theory." It leads—as is demonstrated below—to some unusual consequences, but is able to avoid the discrepancies of quantum theory.

The nonexistence of the microscopic states of macroscopic subsystems of the universe leads to severe difficulties in the interpretation of observation or measurement in terms of information transfer between systems. In particular, since no microscopic state of an organism exists, the principle of "psychophysical parallelism"<sup>(4)</sup> does not apply. In order to understand Eq. (3), the meaning of superpositions of macroscopically different states has to be investigated. Consider, for the moment, a right-handed sugar molecule with wave function  $\varphi_R$ . This is different from an eigenstate of its Hamiltonian  $H_S$ ,  $\varphi_R \pm \varphi_L$ . In contrast to the analogous situation for an ammonia molecule, the tunneling time from  $\varphi_R$  to  $\varphi_L$  is much larger than the age of the universe. The interaction matrix element  $\langle \varphi_R | H_S | \varphi_L \rangle$  is extremely small, as  $H_S$  can at most change the state of two particles. Assume now that an eigenstate  $\varphi_R \pm \varphi_L$  had been prepared. The two components would then interact in different ways with their environment,

$$e^{iHt} \phi(\varphi_R \pm \varphi_L) \approx \phi^{(R)}(t) \varphi_R \pm \phi^{(L)}(t) \varphi_L \equiv \psi^{(R)}(t) \pm \psi^{(L)}(t) \quad (9)$$

(Destruction of the sugar molecule is neglected, and excitations may be taken into  $\phi$ .) With respect to the parity quantum number, the sugar molecule behaves like a macroscopic object—the energy difference between the eigenstates is extremely small. The two world components  $\psi^{(R)}$  and  $\psi^{(L)}$  will behave practically independently after they have been prepared, since  $\langle \psi^{(R)} | H | \psi^{(L)} \rangle$  becomes even smaller with increasing time. There are no transitions between them any more. The "handedness" of the sugar is dynamically stable, whereas one component of the oriented ammonia molecule would emit a photon.

Such a dynamical decoupling of components is even more extreme if  $\varphi_R$  and  $\varphi_L$  represent two states of a pointer corresponding to different positions. Each state will now produce macroscopically correlated states: different images on the retina, different events in the brain, and different reactions of the observer. The different components represent two completely decoupled worlds. This decoupling describes exactly the "reduction of the wave function." As the "other" component cannot be observed any more, it serves only to save the consistency of quantum theory. Omitting this component is justified pragmatically, but leads to the discrepancies discussed above.

This interpretation, corresponding to a "localization of consciousness" not only in space and time, but also in certain Hilbert-space components, has been suggested by Everett<sup>(9)</sup> in connection with the quantization of general relativity, and called the "relative state interpretation" of quantum theory. It amounts to a reformulation of the "psychophysical parallelism" which has in any case become necessary as a consequence of the above discussion of dynamical correlations between states of macroscopic systems.<sup>2</sup> A theory of measurement must necessarily be empty if it does not have a substitute for psychophysical parallelism. Everett's relative state interpretation is ambiguous, however, since the dynamical stability conditions<sup>3</sup> are not considered. This ambiguity is present in the orthodox interpretation of quantum theory as well, where it has always been left to intuition which property of a system is measured "automatically" (e.g., handedness for the sugar, but parity for the ammonia molecule). The dynamical stability appears also to be the cause why microscopic oscillators are observed in energy eigenstates, whereas macroscopic ones occur in "coherent states."<sup>(5)</sup>

According to the twofold localization of consciousness, there are two kinds of subjectivity: The result of a measurement is subjective in that it depends on the world component of the observer; it is objective in the sense that all observers of this world component observe the same result. The question of whether the other components still "exist" after the measurement is as meaningless as asking about the existence of an object while it is not being observed. It is meaningful, however, to ask whether or not the assumption of this existence (i.e., of an objective world) leads to a contradiction.

The probability postulate of quantum theory can be formulated in the following way: Suppose a sequence of equivalent measurements have been performed, each creating an equivalent "branching of the universe." The observer can explain the results by assuming that his final branch has been "chosen randomly" if the components are weighted by their norm. The irreversibility connected with this branching is different from that due to thermodynamical statistics, and thus cannot be explained in terms of the latter. Instead, the effect of branching, i.e., measurement, should be of importance for the foundation of thermodynamics. It seems to be partly taken into account by using the density matrix formalism.<sup>4</sup>

<sup>2</sup>Another suggestion of Wigner's,<sup>(10)</sup> which postulates an active role of consciousness, would require corrections to the equations of motion.

<sup>3</sup>The importance of stability for organic systems has been emphasized by Elsasser.<sup>(11)</sup>

<sup>4</sup>This may indeed be the reason why the foundation of quantum-mechanical thermodynamics appears simpler than that of classical thermodynamics. Proofs of the master equation would, however, be circular again if the process of measurement and hence the density matrix formalism were themselves based on thermodynamics.

The famous paradox of Einstein, Rosen, and Podolski<sup>(12)</sup> is solved straightforwardly: A particle of vanishing spin is assumed to decay into two spin- $\frac{1}{2}$  particles. As a consequence, and according to the axiom of measurement, each particle possesses spin projections of equal probability with respect to any direction in space. After measuring the spin of one particle, however, the spin of the other one is determined. According to Einstein *et al.*, this cannot be true if quantum theory is complete, as there is no interaction with the second particle. The interpretation is that the measurement corresponds to the transformation

$$e^{iHt}\phi(\varphi_1^+\varphi_2^- - \varphi_1^-\varphi_2^+) = \varphi_1^+\varphi_2^-\phi^{(+)}(t) - \varphi_1^-\varphi_2^+\phi^{(-)}(t) \quad (10)$$

where  $\phi^{(+)}$  and  $\phi^{(-)}$  are dynamically decoupled after a short time. Hence, there is one world component in which the experimentalists observe  $\varphi_1^+$  and  $\varphi_2^-$ , another one in which they observe  $\varphi_1^-$  and  $\varphi_2^+$ . As these components cannot "communicate," the result is in accord with the axiom of measurement.

This interpretation of measurement may also explain certain "superselection rules"<sup>(13)</sup> which state, for example, that superpositions of states with different charge cannot occur. It is very plausible that any measurement performed with such a system must necessarily also be a measurement of the charge. Superpositions of states with different charge therefore cannot be observed for similar reasons as those valid for superpositions of macroscopically different states: They cannot be dynamically stable because of the significantly different interaction of their components with their environment, in analogy to the different handedness components of a sugar molecule.

If experimental evidence verifies a spontaneous symmetry-breaking of the vacuum as predicted by many field theories<sup>(14)</sup> this would not prove an asymmetry of the world. One may formally construct invariant wave functions  $\Psi = \int d\Omega U_\alpha \phi$  from symmetry-violating wave functions  $\phi$  (as done for microscopic systems<sup>(15)</sup>). The former cannot be distinguished from its components  $U_\alpha \phi$  if the relative state interpretation is accepted.

It appears that the objective interpretation of quantum theory does not contradict the probability interpretation. It has to be admitted, however, that the "relative state wave function" describes only part of the universe. There is no information on other components except for those which have been created by branching in the past. No estimate can therefore be made on the probability of an inverse branching process, i.e., the spontaneous occurrence of components by accidental overlap.

## ACKNOWLEDGMENT

I wish to thank Prof. E. P. Wigner for encouraging a more detailed version of the third section, and Dr. M. Böhnning for several valuable remarks.

## REFERENCES

1. E. P. Wigner, *Am. J. Phys.* **31**, 6 (1963); B. d'Espagnat, *Nuovo Cimento (Suppl.)* **4**, 828 (1966); T. Earman and A. Shimony, *Nuovo Cimento* **5B**, 332 (1968); J. M. Jauch, E. P. Wigner, and

- M. M. Yanase, *Nuovo Cimento* **48B**, 144 (1967); G. Ludwig, in *Werner Heisenberg und die Physik unserer Zeit* (Braunschweig, 1961).
2. G. Ludwig, *Die Grundlagen der Quantenmechanik* (Berlin, 1954), p. 122 f.; *Z. Physik* **135**, 483 (1953); A. Danieri, A. Loinger, and G. M. Prosperi, *Nucl. Phys.* **33**, 297 (1962); *Nuovo Cimento* **44B**, 119 (1966); L. Rosenfeld, *Prog. Theor. Phys. (Suppl.)* p. 222 (1965); W. Weidlich, *Z. Physik* **205**, 199 (1967).
  3. J. S. Bell, *Rev. Mod. Phys.* **38**, 447 (1966); D. Bohm and J. Bub, *Rev. Mod. Phys.* **38**, 453 (1966).
  4. J. von Neumann, *Mathematische Grundlagen der Quantenmechanik* (Springer, Berlin, 1932) [English translation: Mathematical Foundations of Quantum Mechanics (Princeton University Press, Princeton, N. J., 1955)].
  5. R. J. Glauber, *Phys. Rev.* **131**, 2766 (1963); P. Caruthers and M. M. Nieto, *Am. J. Phys.* **33**, 537 (1965); B. Jancovici and D. Schiff, *Nucl. Phys.* **58**, 678 (1964); C. L. Mehta and E. C. G. Sudarshan, *Phys. Rev.* **138**, B274 (1965).
  6. E. Schrödinger, *Z. Physik* **14**, 664 (1926).
  7. D. Bohm, *Quantum Theory* (Prentice-Hall, Englewood Cliffs, N.J., 1951).
  8. J. M. Jauch, *Helv. Phys. Acta* **33**, 711 (1960).
  9. H. Everett, *Rev. Mod. Phys.* **29**, 454 (1957); J. A. Wheeler, *Rev. Mod. Phys.* **29**, 463 (1957). (
  10. E. P. Wigner, in *The Scientist Speculates*, L. J. Good, ed. (Heinemann, London, 1962), p. 284.
  11. W. M. Elsasser, *The Physical Foundation of Biology* (Pergamon Press, New York and London, 1958).
  12. A. Einstein, N. Rosen, and B. Podolski, *Phys. Rev.* **47**, 777 (1935); D. Bohm and Y. Aharonov, *Phys. Rev.* **108**, 1070 (1957).
  13. G. C. Wick, A. S. Wightman, and E. P. Wigner, *Phys. Rev.* **88**, 101 (1952); E. P. Wigner and M. M. Yanase, *Proc. Natl. Acad. Sci. (US)* **49**, 910 (1963).
  14. W. Heisenberg, *Rev. Mod. Phys.* **29**, 269 (1957); Y. Nambu and G. Jona-Lasinio, *Phys. Rev.* **122**, 345 (1961).
  15. H. D. Zeh, *Z. Physik* **202**, 38 (1967).

# RENORMALIZATION OF MASSLESS YANG-MILLS FIELDS

G.'t HOOFT

*Institute for Theoretical Physics, University of Utrecht,  
Utrecht, The Netherlands*

Received 12 February 1971

**Abstract:** The problem of renormalization of gauge fields is studied. It is observed that the use of non-gauge invariant regulator fields is not excluded provided that in the limit of high regulator mass gauge invariance can be restored by means of a finite number of counter-terms in the Lagrangian. Massless Yang-Mills fields can be treated in this manner, and appear to be renormalizable in the usual sense.

Consistency of the method is proved for diagrams with non-overlapping divergencies by means of gauge invariant regulators, which however, cannot be interpreted in terms of regulator fields. Assuming consistency the  $S$ -matrix is shown to be unitary in any order of the coupling constant. A restriction must be made: no local, parity-changing transformations must be contained in the underlying gauge group. The interactions must conserve parity.

## 1. INTRODUCTION

In recent years the Feynman rules for massless Yang-Mills fields have been established [1–5]. Naive power counting suggests a renormalizable theory; however, in order to carry through a renormalization procedure one must first define a cut-off procedure. And if the cut-off procedure breaks the gauge-invariance of the theory then it is no more clear what the Feynman rules are. The reason is that gauge-invariance, through Ward identities, is essential for the  $S$ -matrix to be unitary.

Thus the problem poses itself as follows: how to find a gauge invariant cut-off procedure. This problem is of course quite the same in quantum electrodynamics. There the problem was solved by Pauli, Villars [6] and Gupta [7] who succeeded in finding a set of regulator fields that could be coupled in a gauge invariant way. Now in the case of massless Yang-Mills fields a gauge invariant regularizing procedure also seems to exist. Unfortunately, however, this procedure cannot be interpreted in terms of fields with indefinite metric and/or wrong statistics, like in the case of electrodynamics. Hence, unitarity and causality are no longer evident.

However, it must be realized that the whole renormalization procedure involves

also the addition of counterterms in the Lagrangian. And in fact the important point is that the total effect of regulator fields and counterterms is to be gauge invariant, at least in the limit of infinite regulator masses. Thus let us suppose now that we have found a set of regulator fields, that makes the various amplitudes finite but destroys the gauge invariance. If we are to restore gauge invariance by means of a finite number of counterterms in the Lagrangian then the gauge-invariance breaking terms in the above mentioned amplitudes must be polynomials of a definite degree in the external momenta, order by order in perturbation theory. But this is precisely the same problem as with the ultra-violet infinities in perturbation theory: the cut-off dependent terms must be polynomials of a definite degree in order for the theory to be renormalizable. Thus the usual proofs of the renormalizability of quantum electrodynamics also guarantee that the unwanted effects of a non-gauge invariant regulator procedure may be off-set by suitably chosen counterterms. Our aim with this procedure is twofold: first, causality is evident, and unitarity can be proven using Cutkosky relations. Secondly, actual calculations are much easier this way, because the counterterms can be fixed easily by applying Ward identities, whereas gauge-invariant regulators become rather complicated particularly at higher orders.

The above point may be illustrated in quantum electrodynamics; in sect. 2 our cut-off procedure is applied to the lowest order photon self energy diagram. Here the unwanted effects of a non-gauge invariant regulator procedure are seen to be such that they can be cured by means of counterterms, one of which has the form of a photon mass term.

One may argue that the method is equivalent with a dispersion relation technique, where the subtraction constants are determined by generalized Ward identities; and that is then sufficient to have a completely gauge invariant theory.

In sect. 3 the situation for massless Yang-Mills fields is outlined. First we use non-gauge invariant regulators, and require that counterterms that remove divergencies are such that Ward identities hold\*.

Consequently, three important questions must be answered:

(i) Do the Ward identities determine the hitherto arbitrary coefficients uniquely? Indeed, we will show that only one arbitrary physical constant remains, being the renormalized coupling constant. Two other arbitrary numbers are unobservable and can be chosen by some convention.

(ii) Are there no internal inconsistencies, like in the PCAC case [8, 9], where no renormalizable counterterm could be found in such a way that PCAC and gauge invariance hold at the same time? In sect. 4 we show a combinatorial proof of the Ward identities, and it appears that many shifts of integration variables are necessary for this proof. Nevertheless, there are no inconsistencies, and for the case of one closed loop we prove this by deriving the gauge invariant set of regulators already referred to (sect. 5). Extension of a similar regulator technique to higher-orders

\* The method of removing infinities by the use of Ward identities and counterterms for Quantum Electrodynamics is described in Jauch and Rohrlich, Theory of photons and electrons, p. 189.

seems possible, but complicated and tricky, and we shall not bother about it in this article.

(iii) Is the resulting  $S$ -matrix unitary? In sect. 5 we generalize the Ward identities, in order to show that the ghost particle intermediate states cancel the intermediate states with non-physically polarized  $W$ -particles. Thus in the unitarity equation only physically (i.e. transversely) polarized  $W$ -particles occur in the intermediate states.

In appendix A a simple formal path integral derivation of the Feynman rules for Yang-Mills fields and the generalized Ward identities is given for both Landau and Feynman gauge. The rules are listed in appendix B.

We use the notation  $k_\mu = (\mathbf{k}, ik_0)$ ;  $k^2 = \mathbf{k}^2 - k_0^2$ . Throughout the paper we confine ourselves to the perturbation expansion. The underlying group here is  $SU(2)$ , though this is not essential. For simplicity also, no other particles with isospin are taken into account, but introducing them does not give rise to any serious difficulty, as long as the matrix  $\gamma^5$  and the tensor  $\epsilon_{\kappa\lambda\mu\nu}$  do not occur in the Lagrangian.

## 2. QUANTUM ELECTRODYNAMICS

In this section we review the situation in quantum electrodynamics. We calculate the contribution of the diagram in fig. 1 to the photon self-energy: a spin  $\frac{1}{2}$  particle forms a closed loop. We do this calculation in order to show the procedure, which can readily be extended to non-Abelian gauge fields.

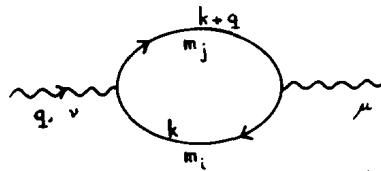


Fig. 1.

The integral diverges quadratically. Now suppose we regularize by replacing the propagator  $(m + i\gamma k)^{-1}$  by

$$\sum_i c_i (m_i + i\gamma k)^{-1}, \quad (2.1)$$

with

$$\sum_i c_i = 0, \quad \sum_i c_i m_i = 0, \quad \sum_i c_i m_i^2 = 0, \quad c_0 = 1, \quad m_0 = m, \quad (2.2)$$

and let ultimately  $m_i$  go to infinity for  $i \neq 0$  ( $c_i$  remain finite).

For finite  $m_i$  the integral now converges and we may shift the integration variable and integrate symmetrically. Then we have

$$\begin{aligned} \Pi_{\mu\nu} &= -\frac{ie^2}{(2\pi)^4} \int d^4k \sum_{ij} c_i c_j \frac{\text{Tr}(m_i - i\gamma k) \gamma_\mu (m_j - i\gamma(k+q)) \gamma_\nu}{(k^2 + m_i^2)((k+q)^2 + m_j^2)} \\ &= -\frac{4ie^2}{(2\pi)^4} \int_0^1 dx \int d^4k \sum_{ij} c_i c_j \frac{(m_i m_j + \frac{1}{2}k^2 - x(1-x)q^2) \delta_{\mu\nu} + 2x(1-x)q_\mu q_\nu}{[k^2 + m_i^2 x + m_j^2(1-x) + q^2 x(1-x)]^2}. \end{aligned} \quad (2.3)$$

Let us define

$$\mu_{ij}^2 \equiv m_i^2 x + m_j^2(1-x) + q^2 x(1-x), \quad (2.4)$$

then we also have

$$\sum_{ij} c_i c_j \mu_{ij}^2 = 0, \quad (2.5)$$

and we can evaluate the convergent integral using

$$\begin{aligned} \int \sum_{ij} c_i c_j \frac{d^4k}{(k^2 + \mu_{ij}^2)^2} &= -i\pi^2 \sum_{ij} c_i c_j \log \mu_{ij}^2, \\ \int \sum_{ij} c_i c_j \frac{m_i m_j d^4k}{(k^2 + \mu_{ij}^2)^2} &= -i\pi^2 \sum_{ij} c_i c_j m_i m_j \log \mu_{ij}^2, \\ \int \sum_{ij} c_i c_j \frac{k^2 d^4k}{(k^2 + \mu_{ij}^2)^2} &= 2i\pi^2 \sum_{ij} c_i c_j \mu_{ij}^2 \log \mu_{ij}^2, \end{aligned} \quad (2.6)$$

so that (2.3) becomes

$$\begin{aligned} \left(\frac{e}{2\pi}\right)^2 \int_0^1 dx \sum_{ij} c_i c_j \{ &\delta_{\mu\nu} (2x(1-x)q^2 + m_i^2 x + m_j^2(1-x) - m_i m_j) \\ &- 2x(1-x)q_\mu q_\nu \} \log [m_i^2 x + m_j^2(1-x) + q^2 x(1-x)]. \end{aligned} \quad (2.7)$$

To see what happens if for  $i \neq 0$   $m_i$  goes to infinity while the  $c_i$  remain finite, we

split off the term  $i = j = 0$  and ignore contributions of order  $q^2/m_i^2$  for  $i \neq 0$ :

$$\begin{aligned} \Pi_{\mu\nu} = & \left(\frac{e}{2\pi}\right)^2 \int_0^1 dx \left\{ 2x(1-x)(q^2\delta_{\mu\nu} - q_\mu q_\nu) \left[ \log(m^2 + q^2x(1-x)) \right. \right. \\ & + \sum_{ij}' c_i c_j \log(m_i^2 x + m_j^2(1-x)) \left. \right] \\ & + \sum_{ij}' c_i c_j \delta_{\mu\nu} (m_i^2 x + m_j^2(1-x) - m_i m_j) \left[ \log(m_i^2 x + m_j^2(1-x)) \right. \\ & \left. \left. + \frac{q^2 x(1-x)}{m_i^2 x + m_j^2(1-x)} \right] + \text{terms } O\left(\frac{q^2}{m_{i \neq 0}^2}\right) \right\} , \end{aligned} \quad (2.8)$$

where  $\Sigma_{ij}'$  denotes the sum over all  $i$  and  $j$  except the term with both  $i = j = 0$ . This result does not satisfy the usual gauge condition

$$q_\mu \Pi_{\mu\nu}(q) = 0 , \quad (2.9)$$

and the renormalized mass of the photon is not evidently zero.

Of course, the reason is that our regulators are not gauge invariant; a vertex where a photon line is attached to particle lines with different masses is not allowed. If we had used Pauli-Vilars-Gupta regulator fields instead of the propagators (2.1), that is, if in formulae (2.3)–(2.8)  $\Sigma_{ij} c_i c_j$  is replaced by

$$\sum_{ij} c_i \delta_{ij} ,$$

then the second term in (2.8) would vanish identically and eq. (2.9) would be fulfilled [6, 7].

However, it is important to note that the gauge non-invariant term in (2.8) is only a polynomial of rank one as a function of  $q^2$ . Let us abbreviate it by

$$\left(\frac{e}{2\pi}\right)^2 (M + L q^2) \delta_{\mu\nu} . \quad (2.10)$$

It can be removed from expression (2.8) if we add a simple counterterm into the Lagrangian\*

$$\Delta \mathcal{L} = -\frac{1}{2} \left(\frac{e}{2\pi}\right)^2 (MA_\mu^2 + L(\partial_\nu A_\mu)^2) . \quad (2.11)$$

\* This implies that terms in the Lagrangian are renormalized, not the fields, as is often done. The difference is merely a scale transformation of the bare quantities.

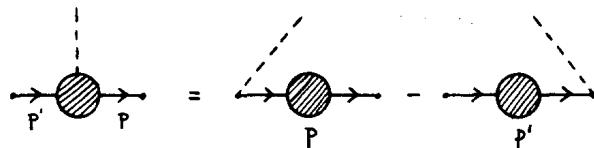
These terms are local and have dimension less than or equal to four, so that causality and renormalizability are not destroyed.

This can be seen to be a very general feature: instead of the gauge invariant Pauli-Villars-Gupta regularization technique we could just as well regularize with the revised propagator (2.1) (which is a non-gauge invariant procedure) and add to the Lagrangian as many local counterterms with dimension less than or equal to four, as desirable. All arbitrary coefficients can then be fixed by requiring the validity of identities like (2.9).

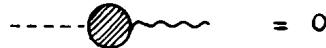
Equations like (2.9) will be called generalized Ward identities\* from now on. They are derived from the usual Ward-Takahashi identity

$$(p' - p)_\mu \Gamma_\mu(p', p) = S_F'^{-1}(p') - S_F'^{-1}(p), \quad (2.12)$$

which can be symbolized as



Here the dashed line denotes a "scalar photon" (a photon line with polarization vector proportional to its own momentum). This identity can be used to derive other equalities for diagrams. For instance



which is precisely eq. (2.9).

In our example we see that the coefficient in front of  $(q^2 \delta_{\mu\nu} - q_\mu q_\nu)$  is still unspecified. This is because we can add freely counter terms proportional to  $F_{\mu\nu} F_{\mu\nu}$  to the Lagrangian because they are gauge invariant themselves. It corresponds to a scale transformation in our definition of the field  $A_\mu$ . So the freedom we have is only a freedom in definition. The most convenient choice is to keep the matrix element of  $A_\mu(x)$  between the vacuum and the one-photon state fixed:

$$\langle 0 | A_\mu(x) | k, \epsilon \rangle = \epsilon_\mu e^{ikx}. \quad (2.13)$$

The renormalized propagator must then have a pole with residue unity at  $k^2 = 0$ , just as the bare propagator.

\* See e.g. J.D.Bjorken and S.D.Drell, Relativistic quantum fields.

So (2.8) must vanish on the mass shell:

$$\Pi_{\mu\nu}(q^2 = 0) = 0 , \quad (2.14)$$

and we derive finally

$$\begin{aligned} \Pi_{\mu\nu}(q^2) &= \left(\frac{e}{2\pi}\right)^2 \int_0^1 dx 2x(1-x)(q^2\delta_{\mu\nu} - q_\mu q_\nu) \\ &\times [\log(m^2 + q^2x(1-x)) - \log m^2] . \end{aligned} \quad (2.15)$$

Once we know that the above mentioned procedure works well, we can go even further and leave the particular set of regulator fields or propagators altogether unspecified. Instead of the identities (2.6) we may use the symbolic expressions:

$$\begin{aligned} \int \frac{d^4k}{(k^2 + \mu^2)^2} &= -i\pi^2 \log \mu^2 + D_1 , \\ \int \frac{k^2 d^4k}{(k^2 + \mu^2)^2} &= 2i\pi^2 \mu^2 \log \mu^2 + D_2 + D_3 \mu^2 , \end{aligned} \quad (2.16)$$

indicating only the terms  $i = j = 0$  in eq. (2.6) explicitly.

The constants  $D_{1,2,3}$  depend on the diagram for which the integral is evaluated, but do not depend on  $\mu$ . Of course, expressions like (2.15) must be handled with great care, but in general they give a very clear idea of where arbitrary numbers enter in the theory. The arbitrariness can only be removed if some additional symmetry property of the system is known, like gauge invariance.

### 3. MASSLESS YANG-MILLS FIELDS

We now consider the Lagrangian of the massless Yang-Mills theory [10]:

$$\mathcal{L}_{YM} = -\frac{1}{4}G_{\mu\nu}G_{\mu\nu} , \quad (3.1)$$

$$G_{\mu\nu}^a = \partial_\mu W_\nu^a - \partial_\nu W_\mu^a + g\epsilon_{abc}W_\mu^b W_\nu^c , \quad (3.2)$$

which is invariant under the local gauge transformation

$$W_\mu^a(x) = f_{ab}(x)W_\mu^b(x) - \frac{1}{2g}\epsilon_{abc}(\partial_\mu f(x)f^{-1}(x))_{cb} . \quad (3.3)$$

If one wants to apply conventional field theory to this model one encounters difficulties [1]. Mandelstam [2] derived Feynman rules for the system using path dependent Green's functions. DeWitt, Faddeev and Popov [3, 4] derived the same rules using a path integral method. We sketch a simple path integral derivation for different gauges in appendix A, and the resulting rules are listed in appendix B:

An auxiliary "ghost particle" appears. In fact it will be seen to cancel the third polarization direction of the W-particles. There is an arbitrariness in gauge, expressed in the parameter  $\lambda$  in the propagator

$$\frac{\delta_{\mu\nu} - \lambda \frac{k_\mu k_\nu}{k^2}}{k^2}$$

Other gauges, like the transversal, can be described in the same way [5].

A path integral derivation of generalized Ward identities is also given in appendix A. A "scalar" W-line



is defined as a W-line with polarization vector  $-ik_\mu$ :

$$\text{---} = -ik_\mu (\text{---}) \quad (3.4)$$

A "transversal line" has a polarization vector  $\epsilon_\mu$  satisfying

$$\begin{aligned} k_\mu \epsilon_\mu &= 0, \\ \epsilon_4 &= 0. \end{aligned} \quad (3.5)$$

A generalized Ward identity is then:

$$\left. \begin{array}{c} \text{on mass shell} \\ \text{transversal} \end{array} \right\} \text{---} \text{---} \text{---} \text{---} \left. \begin{array}{c} \text{on mass shell} \\ \text{transversal} \end{array} \right\} = 0. \quad (3.6)$$

off mass shell

Amplitudes with "longitudinal W-lines" ( $\epsilon_\mu = (-1)^\delta \mu^4 k_\mu$ ) satisfy more complicated Ward identities (cf. sect. 6).

These identities are seen to express the gauge invariance of the theory. For example, the equivalence of the Feynman ( $\lambda = 0$ ) and the Landau gauge ( $\lambda = 1$ ) can be proven using (3.6).

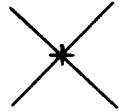
Without much effort one now can verify that the Ward identities are sufficient to prescribe all subtraction constants uniquely, except for the coupling constant. The only needed (and allowed) counterterms are of the following type

$$\text{---} \times \text{---} \delta_{ab} [\delta_{\mu\nu}(C_0 + C_1 k^2) + C_2 k_\mu k_\nu] , \quad (3.7a)$$

$$\text{---} \times \rightarrow \text{---} \delta_{ab} C_3 k^2 , \quad (3.7b)$$



$$-ig C_4 \epsilon_{abc} [\delta_{\beta\gamma}(q-p)_\alpha + \delta_{\gamma\alpha}(k-q)_\beta + \delta_{\alpha\beta}(p-k)_\gamma] , \quad (3.7c)$$



$$\begin{aligned} & -g^2 C_5 [\epsilon_{gac} \epsilon_{gbd} \delta_{\alpha\beta} \delta_{\gamma\delta} + \text{permutations}] \\ & + g^2 C_6 [\delta_{ab} \delta_{cd} (\delta_{\alpha\beta} \delta_{\gamma\delta} + \delta_{\alpha\delta} \delta_{\gamma\beta}) + \text{permutations}] , \end{aligned} \quad (3.7d)$$



$$-ig C_7 q_\alpha , \quad (3.7e)$$

(vertices with more  $\varphi$ -lines do not occur because any amplitude must contain as a factor the momenta of the *outgoing*  $\varphi$ -particles (or ingoing  $\varphi$ -antiparticles) as can be seen from the rules (B.1)–(B.6).

The numbers  $C_1$ ,  $C_3$  and  $C_4$  may be chosen freely, using some convention for the physical amplitude of the  $W$ - and  $\varphi$ -fields, and the definition of the physical coupling constant  $g^{\text{renormalized}}$ . In the Landau gauge moreover,  $C_2$  is immaterial.

According to the Ward identity for the self-energy correction one must have:

$$\text{---} \circ \text{---} + \text{---} \times \text{---} = 0 \quad (3.8)$$

where the counterterm is indicated explicitly, while

$$\text{---} \times \text{---} = \delta_{ab} (C_0 k^2 + C_1 k^4 + C_2 k^4) .$$

So  $C_0$  is fixed and  $C_2$  is expressed in  $C_1$ .

Indeed, an actual calculation of the second-order self energy diagram in the Feynman gauge using the symbolic expressions (2.16) shows:

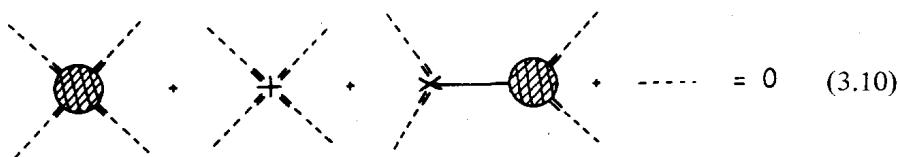
$$\Pi_{\mu\nu}^{ab} = -\frac{g^2}{(4\pi)^2} \delta_{ab} [\frac{10}{3} k^2 \delta_{\mu\nu} - \frac{10}{3} k_\mu k_\nu] \log k^2 + \delta_{ab} [\delta_{\mu\nu} (C_0 + C_1 k^2) + C_2 k_\mu k_\nu] , \quad (3.9)$$

so indeed the Ward identity (3.8) can be satisfied:

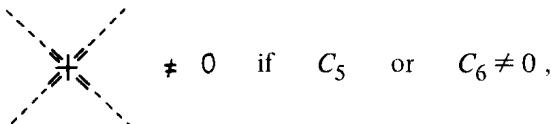
$$C_0 = 0, \quad C_1 + C_2 = 0.$$

The renormalized mass, depending on  $C_0$ , turns out to be zero. Note that the coefficients in front of the terms  $k^2\delta_{\mu\nu} \log k^2$  and  $k_\mu k_\nu \log k^2$  would not be the same if the  $\varphi$ -particle loop had been left out.

For the four-point function we have,



while



so  $C_5$  and  $C_6$  are expressed in terms of the other subtraction constants.

Finally,  $C_7$  can be determined by applying the Ward identity (3.8) for the higher order self-energy diagram of the W-particle, using for instance the BPH procedure of renormalization [11], and the above mentioned observation that

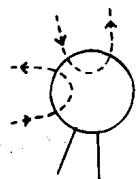
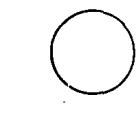


#### 4. COMBINATORIAL PROOF OF THE WARD IDENTITIES

There is no a priori reason why no conflict situation could emerge if we try to satisfy an infinite number of Ward identities using a finite number of counter terms. This problem must be taken seriously, because the algebraic proof of the Ward identities, which will be given below, involves many shifts of integration variables. A proof of the absence of such a conflict will be given only for one closed loop.

Let us introduce some conventions:

stands for the set of diagrams of a given order in  $g$ , and a given number of external transversal W-lines (cf. (3.5)) on mass-shell. There are no longitudinal or scalar external lines. They are denoted explicitly:



stands for the set of diagrams of a given order in  $g$ , and a given number of external transversal W-lines, as above, and in addition a number of external ghost lines and W-lines with arbitrary polarization and momentum, as drawn. The ghost lines are followed inside the graph, which is possible because (B.5) is the only kind of vertex for the ghost particle. The graphs may be disconnected.

The combinatorial proof of the validity of the Ward identities is as follows. From now on we use the Feynman gauge.

Let us perform an infinitesimal gauge transformation in the Lagrangian (3.1):

$$\mathcal{L}_{\text{YM}} = -\frac{1}{4} G_{\mu\nu} G_{\mu\nu} = \mathcal{L}'_{\text{YM}} = -\frac{1}{4} G'_{\mu\nu} G'_{\mu\nu}, \quad (4.1)$$

$$W'_\mu^a(x) = W_\mu^a(x) + g \epsilon_{abc} \Lambda^b(x) W_\mu^c(x) - \partial_\mu \Lambda^a(x), \quad (4.2)$$

$\Lambda$  is some external source which, according to (4.1), remains uncoupled.

Then we must add to all vertices (B.3) and (B.4) all vertices we get from (B.3) and (B.4) if one of the W-lines has been substituted by

$$\begin{array}{ccc} \Lambda \xrightarrow{b} & & \\ \text{W} \xrightarrow{a} & \xrightarrow{c} & -g \epsilon_{abc}, \end{array} \quad (4.3a)$$

$$\begin{array}{ccc} \Lambda \xrightarrow{k} & & \\ \text{W} \xrightarrow{a} & \xrightarrow{c} & -\delta_{ac} i k_\mu. \end{array} \quad (4.3b)$$

(Note: the double line is not meant to be a propagator; (4.3a) is a part of one vertex). Also from the free part of  $\mathcal{L}_{\text{YM}}$  we derive an extra vertex term in  $\mathcal{L}'_{\text{YM}}$ , which appears to be

$$\begin{array}{ccc} \text{W} \xrightarrow{p} & \text{W} \xrightarrow{q} & -g \epsilon_{abc} (\delta_{\mu\nu} p^2 - p_\mu p_\nu - \delta_{\mu\nu} q^2 + q_\mu q_\nu). \end{array} \quad (4.3c)$$

The ghost particle resulting from the use of a certain gauge condition, is not included in our gauge transformation (4.2). Hence, its vertices and propagators are unchanged.

Now it is easy to verify that up to first order in  $\Lambda$  all extra vertices cancel, which

they should do. In diagrams:

$$\begin{array}{c} \text{---} \\ | \\ \text{---} \end{array} + \begin{array}{c} \text{---} \\ | \\ \text{---} \end{array} = 0 \quad (4.4a)$$

$$\begin{array}{c} \nearrow \searrow \\ \text{---} \end{array} + \begin{array}{c} \nearrow \searrow \\ \text{---} \end{array} + \begin{array}{c} \nearrow \searrow \\ \text{---} \end{array} + \begin{array}{c} \nearrow \searrow \\ \text{---} \end{array} = 0 \quad (4.4b)$$

$$\begin{array}{c} \nearrow \searrow \\ \text{---} \end{array} + \begin{array}{c} \nearrow \searrow \\ \text{---} \end{array} + \begin{array}{c} \nearrow \searrow \\ \text{---} \end{array} + \begin{array}{c} \nearrow \searrow \\ \text{---} \end{array} = 0 \quad (4.4c)$$

(4.3b) is of the type which occurs in our Ward identities. We now see that it can be replaced by (4.3a) and (4.3c) using eqs. (4.4), except for the connections with the ghost particle. So as

$$\begin{array}{c} \text{---} \\ | \\ \text{---} \end{array} = - \begin{array}{c} \text{---} \\ | \\ \text{---} \end{array} + \frac{1}{2} \begin{array}{c} \text{---} \\ | \\ \text{---} \end{array} + \frac{1}{6} \begin{array}{c} \text{---} \\ | \\ \text{---} \end{array} \quad (4.5)$$

(Note the explicitly written minus sign for the  $\varphi$ -loop and the combinatory factors, because the blobs are already symmetrized) we have

$$(4.5) = - \begin{array}{c} \text{---} \\ | \\ \text{---} \end{array} - \frac{1}{2} \begin{array}{c} \text{---} \\ | \\ \text{---} \end{array} - \frac{1}{2} \begin{array}{c} \text{---} \\ | \\ \text{---} \end{array} - \frac{1}{6} \underbrace{\begin{array}{c} \text{---} \\ | \\ \text{---} \end{array}}_C \quad (4.6)$$

A                    B                    C

eq. (4.6) may be written as

$$\begin{array}{ccccccc}
 \text{Diagram} & = & \text{Diagram A} & - \frac{1}{2} \text{Diagram C}_1 & - \frac{1}{2} \text{Diagram B} & - \frac{1}{2} \text{Diagram C}_2 & - \text{Diagram C}_3 \\
 & & \text{A} & \text{C}_1 & \text{B} & \text{C}_2 & \text{C}_3
 \end{array} \quad (4.7)$$

(Of course,  $C_1$  equals  $C_2$ .)

Note that  $C_3$  cancels those diagrams contained in  $C_1$  and  $C_2$  where the double line is attached to a ghost vertex.

The next step is a propagator identity which is related to invariance of the gauge condition under special gauge transformations  $\Lambda$  with  $\partial_\mu(\partial_\mu\Lambda^a + g\epsilon_{abc}W_\mu^b\Lambda^c) = 0$ :

$$\begin{array}{ccccccc}
 \text{Diagram} & + & \text{Diagram P} & + & \text{Diagram P} & + & \text{Diagram P} = 0 \\
 & & \text{P} & & \text{P} & & \text{P}
 \end{array} \quad (4.8a)$$

$$\begin{array}{ccccccc}
 \text{Diagram} & + & \text{Diagram P} & + & \text{Diagram P} & = & 0 \\
 & & \text{P} & & \text{P} & & \text{P}
 \end{array} \quad (4.8b)$$

The  $P$  denotes a transversal  $W$ -line on mass shell (cf. (3.5)). Note again that the double line is no propagator.

Eq. (4.8a) is the Yang-Mills counterpart of the usual Ward-Takahashi identity (2.12) for bare electron propagators and vertex functions. In the last two terms the dashed line (" $\Lambda$ -line") has the same vertices and propagators as the ghost particle (" $\varphi$ -line", compare (B.2) and (B.5)). If some of the lines in (4.8) are parts of a closed loop these identities are true provided one may shift integration variables. This is the reason why subtraction constants must be chosen carefully.

Applying eqs. (4.8) to eq. (4.7) we find

$$\begin{array}{ccccccc}
 \text{Diagram} & = & \text{Diagram A} & - & \text{Diagram C}_1 & - & \text{Diagram C}_2 \\
 & & \text{A} & & \text{C}_1 & & \text{C}_2
 \end{array} \quad (4.9)$$

Eq. (4.9) can now be iterated, but then we must include the possibility that the

$\Lambda$ -line forms a closed loop and is attached to itself. The result is:

$$(4.10)$$

Using one more identity

$$(4.11)$$

we have

$$(4.12)$$

Substituting (4.3b) into (4.3a) one obtains another vertex, for which the following equation holds:

$$(4.13)$$

Consequently the derivation remains valid even if there are more off-mass shell scalar  $W$ -lines:

$$(4.14a)$$

which is the graphical notation for the formula

$$\frac{\partial}{\partial x_{\mu_1}^1} \dots \frac{\partial}{\partial x_{\mu_N}^N} \langle \text{out} | T^*(W_{\mu_1}^{a_1}(x^1) \dots W_{\mu_N}^{a_N}(x^N)) | \text{in} \rangle = 0 , \quad (4.14b)$$

in conventional field theory.

From this algebraic derivation of the Ward identities we draw the following conclusion: if we succeed in regularizing graphs containing one of the auxiliary vertices (4.3a)–(4.3c) in such a way that eqs. (4.8), (4.11) and (4.13) remain valid *also inside closed loops*, then we acquire gauge invariant amplitudes (amplitudes satisfying (4.14)).

## 5. GAUGE INVARIANT REGULATORS

In this section we construct a set of regulators satisfying all requirements formulated in the previous section, but we confine ourselves to the one closed-loop case. The mere existence of these regulators implies that no conflict situation arises if one uses Ward identities for calculating subtraction constants in the first quantum-mechanical correction, instead of gauge invariant regulators.

The procedure is as follows. Note that the identities (4.8), (4.11) and (4.13) are not only valid in a four-dimensional Minkowsky space, but we may add another dimension. Then the momenta  $k_\mu$  have five components, and the fields  $W_\mu^a$  have 15 components. Let for all diagrams with one closed loop the external momenta be in the Minkowsky space, that is, only their first four components differ from zero. Let the momenta inside the closed loop have one more component of fixed length  $M$  in a fixed fifth direction. Because of conservation of momentum,  $M$  is the same for all propagators of the closed loop. With this interpretation in mind, we may now reformulate the Feynman rules, which now contain an extra parameter  $M$ . Furthermore, they depend on which of the propagators belong to the closed loop; those propagators will be denoted by a \*.

The  $W$ - and  $\varphi$ -propagators inside the closed loop are replaced by:



$$\frac{\delta_{ab}\delta_{\mu\nu}}{k^2 + M^2} , \quad (5.1a)$$



$$\frac{\delta_{ab}}{k^2 + M^2} . \quad (5.1b)$$

The vertices (B.3)–(B.5) remain the same, as well as the propagators (B.1) and (B.2) in the tree parts of a graph. In (5.1a) we let the indices  $\mu, \nu$  run from 1 to 4 as usual. The fifth polarization direction of the  $W$ -field is treated as a new particle,

which only occurs inside the closed loop:

$$\begin{array}{c} * \\ \vdots \vdots \vdots \vdots \vdots \vdots \vdots \end{array} \quad \frac{\delta_{ab}}{k^2 + M^2}. \quad (5.1c)$$

It has the vertices:

$$\begin{array}{c} * \\ \vdots \vdots \vdots \vdots \vdots \vdots \vdots \end{array} \quad Mg \epsilon_{abc} \delta_{\alpha\gamma}, \quad (5.1d)$$

$$\begin{array}{c} * \\ \vdots \vdots \vdots \vdots \vdots \vdots \vdots \end{array} \quad -Mg \epsilon_{abc}, \quad (5.1e)$$

(note that the factors  $\pm i$  at each end of a crossed line have cancelled), and

$$\begin{array}{c} * \\ \vdots \vdots \vdots \vdots \vdots \vdots \vdots \end{array} \quad -ig \epsilon_{abc} (q-p)_\alpha, \quad (5.1f)$$

$$\begin{array}{c} * \\ \vdots \vdots \vdots \vdots \vdots \vdots \vdots \end{array} \quad -g^2 (\epsilon_{gac} \epsilon_{gbd} + \epsilon_{gad} \epsilon_{gbc}) \delta_{\alpha\beta}. \quad (5.1g)$$

Now with vertices (5.1f) and (5.1g) one may have closed loops of crossed lines, but these contributions are gauge invariant themselves, since the vertices (5.1f) and (5.1g) are precisely those of an ordinary isospin one scalar particle. So we may exclude diagrams with closed loops of crossed lines without invalidating the Ward identities. The above vertices with the rule of no closed loop of crossed lines define a set of diagrams which, up to one loop, satisfy the Ward identities. For  $M = 0$  we have the diagrams of the massless theory. For  $M$  non-zero we have diagrams that may be used as regulator diagrams.

Consider now the sum of diagrams of the massless theory and regulator diagrams. Choosing the appropriate integration variables (remember that each individual contribution may be infinite, and relative shifts of integration variables may give different results) and furthermore regulators with masses  $M_i$  and signs  $e_i$ , in such a way that

$$\begin{aligned} \sum e_i &= 0, & e_0 &= 1, \\ \sum e_i M_i^2 &= 0, & M_0 &= 0, \end{aligned} \quad (5.2)$$

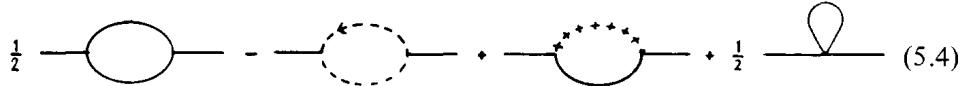
we obtain a finite result.

One may choose convenient, finite values for

$$\sum_{i \neq 0} e_i \log M_i^2 = -A, \quad \sum e_i M_i^2 \log M_i^2 = B. \quad (5.3)$$

In the limit  $M_{i \neq 0} \rightarrow \infty$  we find the desired gauge invariant amplitudes.

Let us demonstrate this regulator technique for the second order self-energy contributions to the  $W$ -propagator:



Using expressions (2.6) we find

$$\begin{aligned} \Pi_{\mu\nu}^{ab}(k) &= \frac{-g^2}{(4\pi)^2} \delta_{ab} \int_0^1 dx \sum_i e_i [ \{ k^2 (5 - 10x(1-x)) \delta_{\mu\nu} \\ &\quad - k_\mu k_\nu (2 + 8x(1-x)) \} \log(M_i^2 + x(1-x)k^2) \\ &\quad - 6M_i^2 \delta_{\mu\nu} \log(M_i^2 + x(1-x)k^2) + 6M_i^2 \delta_{\mu\nu} \log M_i^2 ] . \end{aligned} \quad (5.5)$$

Indeed, one may convince oneself that this satisfies the Ward identity

$$k_\mu k_\nu \Pi_{\mu\nu}^{ab}(k) = 0. \quad (5.6)$$

In the limit  $M_{i \neq 0} \rightarrow \infty$  we have

$$\Pi_{\mu\nu}^{ab} = -\frac{g^2}{(4\pi)^2} \delta_{ab} (k^2 \delta_{\mu\nu} - k_\mu k_\nu) [\frac{10}{3} \log k^2 - \frac{10}{3} A - \frac{56}{9}]. \quad (5.7)$$

The number  $A$  is the logarithm of a suitably chosen reference mass. It must have the same value for all graphs with one closed loop.

It must be emphasized that even if our regulator method appears very similar to the Pauli-Villars method it is in fact very different. The regulators do not correspond to fields in Lagrangians etc., and the procedure works only for one closed loop. In fact the above is just a convenient way of implementing the scheme pro-

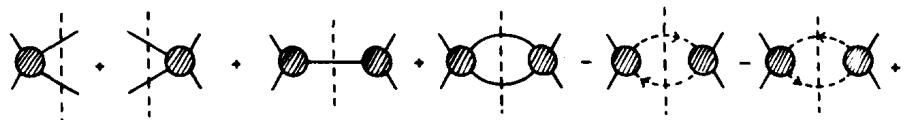
posed in the beginning. Tentative investigation shows that probably a modification of this regulator technique can produce finite gauge invariant amplitudes at higher orders. As yet we shall consider this as a conjecture. It is important to note that this technique of introducing more dimensions only works if the matrix  $\gamma^5$  and the tensor  $\epsilon_{\kappa\lambda\mu\nu}$  do not occur in the Lagrangian.

## 6. UNITARITY

In proving unitarity of the  $S$ -matrix one has to deal with on mass-shell amplitudes. We are then confronted with infrared difficulties. Now if we add a very small mass term  $\kappa^2$  in the propagators, then the on mass-shell amplitudes (in finite order of  $g$ ) are proportional to some power of  $\log \kappa^2$ . The Ward identities however, are violated with terms proportional to  $\kappa^2, \kappa^2 \log \kappa^2$ , etc. So we can still use these Ward identities keeping  $\log \kappa^2$  finite, but ignoring terms proportional to  $\kappa^2, \kappa^2 \log \kappa^2$  etc. For instance, in the regularized expressions in sect. 5 we might put  $M_0 = \kappa \neq 0$ , but ignore the crossed line with mass  $\kappa$ , because it is coupled with strength  $\kappa^2$ .

We shall not go into the problems of the physical interpretation of these infrared divergencies.

To compute imaginary parts we shall make use of the well-known Cutkosky rules [12] :



$$+ (\text{graphs with more than two lines cut through}) = 0, \quad (6.1)$$

where at the right-hand side of the dashed line the  $i\epsilon$  in the propagators is replaced by  $-i\epsilon$ , and an extra minus sign is introduced for each propagator and each vertex. The blobs are at least of order one in  $g$ . Now, if in the blobs of (6.1) all graphs are added, including disconnected ones, such that the total order in  $g$  is kept fixed, then equation (6.1) is an identity, whatever the choice of our subtraction coefficients may be, provided that we use the following rules:

$$2\pi\delta(k^2)\theta(k_0)\delta_{ab}, \quad (6.2)$$

$$\left. \begin{array}{c} \text{Diagram 1: } \text{horizontal line } k, \text{ dashed line } k, \text{ external } a, b \\ \text{Diagram 2: } \text{dashed line } k, \text{ horizontal line } k, \text{ external } a, b \\ \text{Diagram 3: } \text{dashed line } k, \text{ horizontal line } k, \text{ external } a, b \end{array} \right\} 2\pi\delta(k^2)\theta(k_0^2)\delta_{ab}, \quad (6.3)$$

(a dashed line going through an external particle-line has no special meaning, except that it separates the ingoing lines from the outgoing lines).

Now if we can prove a slightly different equation,

$$\dots + \dots + \dots + \dots = 0 \quad (6.4)$$

with

$$\text{standing for } 2\pi\delta(k^2)\theta(k_0)\delta_{ab}\left(\delta_{\mu\nu}-\frac{k_\mu k_\nu}{|k|^2}\right)(1-\delta_{\mu 4})(1-\delta_{\nu 4}), \quad (6.5)$$

then unitarity has been proven, for the case that bosons with a given isospin have only two helicity states, like the photons. We shall prove eq. (6.4) from eq. (6.1) provided that we only look at the transverse components of the other outgoing lines. Let us first consider the case of only two intermediate particles. Define

$$2\pi\delta(k^2)\theta(k_0)\delta_{ab}\frac{-i\bar{k}_\nu}{2|k|^2}, \quad \bar{k}_\nu \equiv (-1)^{\delta_{\nu 4}} k_\nu \quad (6.6)$$

$$2\pi\delta(k^2)\theta(k_0)\delta_{ab}\frac{i\bar{k}_\nu}{2|k|^2}.$$

A useful equation is:

$$\delta_{\mu\nu} = \frac{1}{2} \frac{k_\mu \bar{k}_\nu + \bar{k}_\mu k_\nu}{|k|^2} + \left( \delta_{\mu\nu} - \frac{k_\mu k_\nu}{|k|^2} \right) (1 - \delta_{\mu 4})(1 - \delta_{\nu 4}) \quad \text{if } k^2 = 0. \quad (6.7)$$

Symbolically:

$$(6.8)$$

Also we have

$$(6.9)$$

We shall apply the Ward identities

$$(6.10)$$

Moreover, we need a generalization of the Ward identities (4.14) for amplitudes with on mass-shell ghost particles and non-physically polarized W-particles, in particular W-particles with polarization vector  $e_\mu$  not satisfying  $k_\mu e_\mu = 0$ . Formula (4.8b) is extended to

$$(6.11a)$$

where the arrow in  $\dashrightarrow \mu$  stands for multiplication with  $-ik_\mu$ , and the lines with a circle are taken on mass-shell ( $k^2 = 0$ ). Note that the last graph in (6.11a) vanishes if multiplied with a transversal polarization vector  $e_\mu$ . We have also

$$(6.11b)$$

Applying again the combinatorics of sect. 4 we derive the generalized Ward identity

$$\text{Diagram showing two configurations of a particle with spin and position. The first configuration has a dashed arrow pointing up from the center. The second configuration has a dashed arrow pointing down from the center. An equals sign is between them.} \quad (6.12)$$

(This identity is not altered if other gauge invariant interactions are introduced. The other isospin particles must then be on mass-shell).

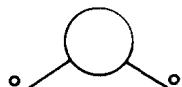
Equipped with eqs. (6.8), (6.9), (6.10) and (6.12) we derive

$$\begin{aligned}
 & \text{Diagram 1} = \text{Diagram 2} \\
 & = \text{Diagram 3} - \text{Diagram 4} - \text{Diagram 5} \\
 & = \text{Diagram 6} - \text{Diagram 7} - \text{Diagram 8}
 \end{aligned} \tag{6.13}$$

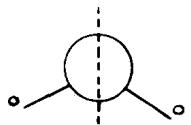
from which eq. (6.4) follows, as long as we confine ourselves to the contributions with at most two particles in the intermediate states.

In the same way it can be shown for intermediate states with more than two particles that the ghost particles cancel the non-physical polarization directions of the W-bosons. In principle this can be verified by writing down further generalizations of the Ward identity (6.12), but a more straightforward proof of this cancellation goes as follows. We apply induction with respect to the number of particles in the intermediate states.

Suppose we have a diagram



(the external lines being on mass-shell). Let then



stand for the sum of all graphs acquired by cutting the former diagram in all possible ways, except that at least one vertex must remain at either side of the dashed line.

Applying again the Cutkosky rule to the left-hand side of (6.12):

$$\begin{array}{c} \text{Diagram 1} \\ + \\ \text{Diagram 2} \\ + \\ \text{Diagram 3} \end{array} = 0 \quad (6.14)$$

one derives easily:

$$\begin{array}{c} \text{Diagram 1} \\ = \\ \text{Diagram 2} \end{array} \quad (6.15)$$

with the external lines on mass shell.

Now careful examination of the underlying propagator identities and combinatorics leads to the observation that eq. (6.15) is also valid if the total number of cut propagators is kept fixed at both sides. So if we introduce the notation

$$\begin{array}{c} \text{Diagram 1} \\ = \\ \text{Diagram 2} \end{array} \quad (6.16)$$

$N$  denoting the total number of cut propagators, then (6.15) reads:

$$\begin{array}{c} \text{Diagram 1} \\ = \\ \text{Diagram 2} \end{array} \quad (6.17)$$

for all  $N$ . Moreover, one can impose the restriction that the cutting line must pass through both of the explicitly denoted external lines in (6.17), and then we get:

$$\begin{array}{c} \text{Diagram 1} \\ = \\ \text{Diagram 2} \end{array} \quad (6.18)$$

Now suppose that for a certain value of  $N$

$$\begin{array}{c} \text{Diagram 1} \\ = \\ \text{Diagram 2} \end{array} \quad (6.19)$$

then we have

$$\begin{aligned}
 & \text{Diagram with } N+1 \text{ vertices} \\
 & = \text{Diagram with } N \text{ vertices} + \text{Diagram with } N+1 \text{ vertices} \\
 & = \text{Diagram with } N \text{ vertices} - \text{Diagram with } N \text{ vertices} - \text{Diagram with } N \text{ vertices} \\
 & = \text{Diagram with } N \text{ vertices} - \text{Diagram with } N \text{ vertices} - \text{Diagram with } N \text{ vertices} \\
 & = \text{Diagram with } N+1 \text{ vertices}
 \end{aligned} \tag{6.20}$$

which completes the proof by induction.

So the  $S$ -matrix is unitary in a Hilbert space with only plane wave W-particle states, in which each particle has helicity  $\pm 1$ . A necessary condition is that subtraction constants are chosen in such a way that all generalized Ward identities are satisfied.

## 7. CONCLUSION

Massless YM fields can be renormalized. A formal regulator procedure exists, at least for diagrams with one closed loop, but the simplest way to deal with the divergencies is to use the subtracted expressions (2.16) for divergent integrals, calculating subtraction constants by means of the Ward identities. In this article we have not gone into the details of a regulator technique for diagrams with more loops, so as yet a consistency proof of the Ward identity method for removing overlapping divergencies, is lacking.

With this restriction, we have proven that the resulting  $S$ -matrix is unitary, if infrared divergencies are dealt with in a proper way. There is only one physical parameter in the theory, which is the coupling constant  $g$ . The renormalized mass of the bosons is zero (at least, in perturbation theory).

The author is greatly indebted to Prof. M.Veltman for many helpful discussions and critical remarks.

## APPENDIX A

*Path integral derivation of Feynman rules for massless Yang-Mills fields*

The Feynman path integral expression for the amplitude is

$$\langle \text{out} | \text{in} \rangle = \int \prod_{x,\mu,a} dW_\mu^a(x) \exp \{ iS_{\text{YM}}[W] \}$$

$$\equiv \int \mathcal{D}W \exp iS_{\text{YM}}[W] , \quad (\text{A.1})$$

where  $a$  denotes isospin,  $\mu$  the Lorentz vector component, and  $S_{\text{YM}}[W] = \int \mathcal{L}_{\text{YM}}(x) dx$  is the (unrenormalized) Yang-Mills action functional. Now if the asymptotic states are invariant under local gauge transformations  $\Omega$ , that is

$$\Omega |\text{in}\rangle = |\text{in}\rangle , \quad \Omega |\text{out}\rangle = |\text{out}\rangle ,$$

then the integrand, as well as the measure  $\mathcal{D}W$ , are invariant under local gauge transformations.

In order to extract the infinite constant arising from this invariance we alter expression (A.1) by multiplying with a delta-function  $\delta(\log \Omega)$  (defined in terms of the same measure  $\mathcal{D}W$ ) where  $\Omega$  is defined such, that the field

$$W' = \Omega^{-1}(W)$$

satisfies a special gauge condition. We choose the gauge

$$\partial_\mu W'_\mu^a(x) = C^a(x) , \quad (\text{A.2})$$

with  $C^a(x)$  a fixed function. Then expression (A.1) becomes

$$\begin{aligned} \int \mathcal{D}W \delta(\log \Omega) \exp iS_{\text{YM}}[W] &= \int \mathcal{D}W \delta(\partial_\mu W_\mu^a - C^a) \\ &\times \det \left( \frac{\partial}{\partial \Omega(x')} \partial_\mu W_\mu^a(x) \right) \exp iS_{\text{YM}}[W] \end{aligned} \quad (\text{A.3})$$

In order to calculate the determinant we only need to know the change of  $\partial_\mu W_\mu^a(x)$  under an infinitesimal gauge transformation  $\Lambda^b(x)$ :

$$\begin{aligned} \partial_\mu W_\mu^{a'} &= \partial_\mu W_\mu^a + \epsilon_{abc} \partial_\mu (\Lambda^b W_\mu^c) - g^{-1} \partial^2 \Lambda^a \\ &= \partial_\mu W_\mu^a - g^{-1} \partial_\mu (D_\mu \Lambda)^a \end{aligned} \quad (\text{A.4})$$

( $D_\mu$  is the covariant derivative and  $g$  is the coupling constant).

So we must calculate the determinant of the operator  $g^{-1}\partial_\mu D_\mu$ . This we do with the following trick. Note that even for a non-hermitean matrix  $A_{ij}$  the identity

$$\frac{1}{\det A} = C \int \prod_i d \operatorname{Re} z_i d \operatorname{Im} z_i \exp i(z^*, Az) \quad (\text{A.5})$$

holds, where  $C$  is a trivial constant. So we write in a symbolic notation eq. (A.3) as

$$\int \mathcal{D}W \delta(\partial_\mu W_\mu^\alpha - C^\alpha) \int \mathcal{D}'\varphi \exp \{iS_{\text{YM}}[W] + i \int \varphi^*(x) \partial_\mu D_\mu \varphi(x) dx\}. \quad (\text{A.6})$$

$\varphi^\alpha(x)$  is a complex scalar particle field. The notation is symbolic because the determinant in eq. (A.3) stands in the numerator and not in the denominator like in eq. (A.5). But this only means that we have to add a factor  $-1$  for each closed loop of  $\varphi$ 's, as can easily be established. It is denoted by the prime in  $\mathcal{D}'\varphi$ .

If  $C^\alpha$  is put equal to zero, we get the rules derived by Faddeev and Popov [4]. The transversal propagators

$$\delta_{ab} \frac{\delta_{\mu\nu} - \frac{k_\mu k_\nu}{k^2}}{k^2}, \quad (\text{A.7})$$

emerge (Landau gauge)<sup>†</sup>. We can get rid of the annoying  $k_\mu k_\nu$  term by noting that expression (A.6) is completely independent of the choice of  $C^\alpha(x)$ . So we may integrate over all values of  $C$ , together with an arbitrary weight function  $\exp iS'[C]$ .

We then get

$$\int \mathcal{D}W \int \mathcal{D}'\varphi \exp \{iS_{\text{YM}}[W] - i \int (\partial_\mu \varphi)^* D_\mu \varphi dx + iS'[\partial_\mu W_\mu]\}. \quad (\text{A.8})$$

$S'[\partial_\mu W_\mu]$  may be chosen such that it cancels the corresponding term in  $S_{\text{YM}}[W]$  and we then find the Feynman gauge, with propagators

$$\frac{\delta_{ab} \delta_{\mu\nu}}{k^2}. \quad (\text{A.9})$$

The resulting Feynman rules are listed in appendix B.

#### Ward identities

We first derive Ward identities in the Landau gauge. Let us treat  $C^\alpha$  in expression (A.6) as a source function and make an expansion with respect to it. Even with out-

<sup>†</sup> The  $i\epsilon$  in a propagator is not found by the path integral method. Its sign is dictated by unitarity and is essential for derivation of the Cutkosky rules (sect. 6).

or ingoing particles at plus or minus infinity expression (A.6) is independent of  $C^a$ . So all expansion terms with respect to  $C^a$  must be zero except the first.

In order to derive the Feynman rules for the expansion terms we must treat the transversal and longitudinal parts of the  $W$ -field separately. Integration over the transversal part leads to the Feynman rules (B.1)–(B.6), with  $\lambda = 1$ , but the fact that  $\partial_\mu W_\mu$  now is  $C$  and not zero gives us the additional  $C$ -lines:

$$\text{---} \overset{\mathbf{k}}{\times} \text{---} \overset{\mathbf{C}}{\times} \quad \delta_{ab} \frac{-ik_\mu}{k^2} e^{-ikx}, \quad (\text{A.10})$$

where the cross denotes the action of the “source”  $C^b(x)$ , and the double line simply acts as a normal Yang-Mills boson. (The derivation is done by making  $\partial_\mu W_\mu$  variable and adding  $-\alpha(\partial_\mu W_\mu^a - C^a)^2$  to the Lagrangian, which gives rise to a delta function for  $\alpha \rightarrow \infty$ .)

We can now formulate our Ward identity in the Landau gauge: *The total contribution of all diagrams with a given (non-zero) number of  $C$ -lines off mass-shell, and a given number of in- or outgoing lines on mass-shell, is zero.*

This rule is visualized in the diagram notation (3.6), and corresponds to formula (4.14b).

Eq. (3.6) greatly resembles the corresponding Ward identities in quantum electrodynamics, the only difference being that we have to contract *all* off mass-shell lines with their own momentum (that is, choose a polarization vector proportional to their own momentum). The outgoing lines must be physical, that is, their polarization vector must be orthogonal to their own momentum.

In the Feynman gauge we can do something similar. In expression (A.8) we made the choice

$$S'[C] = \int dx \{-\frac{1}{2}C^2(x)\}.$$

Now we add a source function  $J(x)$ :

$$S'[C] = \int dx \{-\frac{1}{2}(C(x) - J(x))^2\}. \quad (\text{A.11})$$

Again, the result must be independent of  $J(x)$ .

The Feynman rules are those of appendix B, with  $\lambda = 0$ , together with a  $J$ -source contribution which is the same as (A.10) except for the (immaterial) factor  $1/k^2$ . So the Ward identities in this case are again those of eqs. (3.6) and (4.14).

## APPENDIX B

Feynman rules for massless Yang-Mills fields

$$W: \quad \begin{array}{c} a \quad b \\ \mu \quad \nu \\ \hline \end{array} \quad \frac{\delta_{ab}}{k^2 - i\epsilon} \left( \delta_{\mu\nu} - \lambda \frac{k_\mu k_\nu}{k^2 - i\epsilon} \right) \quad \begin{array}{l} \lambda = 1 \text{ Landau gauge}, \\ \lambda = 0 \text{ Feynman gauge}, \end{array} \quad (B.1)$$

$$\phi: \quad \begin{array}{c} a \quad b \\ \alpha \quad \beta \\ \hline \end{array} \quad \frac{\delta_{ab}}{k^2 - i\epsilon}, \quad (B.2)$$

$$\begin{array}{c} a, \alpha, k \\ \backslash \quad / \\ b, \beta, p \quad c, \gamma, q \\ \hline \end{array} \quad -ig \epsilon_{abc} [\delta_{\beta\gamma}(q-p)_\alpha + \delta_{\gamma\alpha}(k-q)_\beta + \delta_{\alpha\beta}(p-k)_\gamma], \quad (B.3)$$

$$\begin{array}{c} a, \alpha \quad b, \beta \\ \diagdown \quad \diagup \\ c, \gamma \quad d, \delta \\ \diagup \quad \diagdown \end{array} \quad \begin{aligned} & -g^2 \epsilon_{gac} \epsilon_{gbd} (\delta_{\alpha\beta} \delta_{\gamma\delta} - \delta_{\alpha\delta} \delta_{\gamma\beta}) \\ & -g^2 \epsilon_{gad} \epsilon_{gbc} (\delta_{\alpha\beta} \delta_{\delta\gamma} - \delta_{\alpha\gamma} \delta_{\delta\beta}) \\ & -g^2 \epsilon_{gab} \epsilon_{gcd} (\delta_{\alpha\gamma} \delta_{\beta\delta} - \delta_{\alpha\delta} \delta_{\beta\gamma}), \end{aligned} \quad (B.4)$$

$$\begin{array}{c} a, \alpha \\ \backslash \quad / \\ b, \beta \quad c, \gamma \\ \diagup \quad \diagdown \end{array} \quad -ig \epsilon_{abc} q_\alpha, \quad (B.5)$$

(at the vertices all momenta are defined to be inwards).

$$\text{For each closed loop of } \varphi \text{ particles: } -1. \quad (B.6)$$

As usual: a factor  $1/(2\pi)^4 i$  for each propagator and  $(2\pi)^4 i$  for each vertex.

## REFERENCES

- [1] R.P.Feynman, Acta Phys. Polon. 24 (1963) 697.
- [2] S.Mandelstam, Phys. Rev. 175 (1968) 1580; 1604.
- [3] B.S.DeWitt, Phys. Rev. 162 (1967) 1195; 1239.
- [4] L.D.Faddeev and V.N.Popov, Phys. Letters 25B (1967) 29.
- [5] E.S.Fradkin and I.V.Tyutin, Phys. Rev. D2 (1970) 2841.
- [6] W.Pauli and F.Villars, Rev. Mod. Phys. 21 (1949) 434.
- [7] S.N.Gupta, Proc. Phys. Soc. 66 (1953) 129.
- [8] J.S.Bell and R.Jackiw, Nuovo Cimento 60A (1969) 47.
- [9] St.L.Adler, Phys. Rev. 177 (1969) 2426.
- [10] C.N.Yang and R.L.Mills, Phys. Rev. 96 (1954) 191.
- [11] K.Hepp, Comm. Math. Phys. 2 (1966) 301.
- [12] R.E.Cutkosky, J. Math. Phys. 1 (1960) 429;  
M.Veltman, Physica 29 (1963) 186.



## Black Holes and Entropy\*

Jacob D. Bekenstein†

*Joseph Henry Laboratories, Princeton University, Princeton, New Jersey 08540  
and Center for Relativity Theory, The University of Texas at Austin, Austin, Texas 78712‡  
(Received 2 November 1972)*

There are a number of similarities between black-hole physics and thermodynamics. Most striking is the similarity in the behaviors of black-hole area and of entropy: Both quantities tend to increase irreversibly. In this paper we make this similarity the basis of a thermodynamic approach to black-hole physics. After a brief review of the elements of the theory of information, we discuss black-hole physics from the point of view of information theory. We show that it is natural to introduce the concept of black-hole entropy as the measure of information about a black-hole interior which is inaccessible to an exterior observer. Considerations of simplicity and consistency, and dimensional arguments indicate that the black-hole entropy is equal to the ratio of the black-hole area to the square of the Planck length times a dimensionless constant of order unity. A different approach making use of the specific properties of Kerr black holes and of concepts from information theory leads to the same conclusion, and suggests a definite value for the constant. The physical content of the concept of black-hole entropy derives from the following generalized version of the second law: When common entropy goes down a black hole, the common entropy in the black-hole exterior plus the black-hole entropy never decreases. The validity of this version of the second law is supported by an argument from information theory as well as by several examples.

### I. INTRODUCTION

A black hole<sup>1</sup> exhibits a remarkable tendency to increase its horizon surface area when undergoing any transformation. This was first noticed by Floyd and Penrose<sup>2</sup> in an example of the extraction of energy from a Kerr black hole by means of what has come to be known as a Penrose process.<sup>3</sup> They suggested that an increase in area might be a general feature of black-hole transformations. Independently, Christodoulou<sup>4,5</sup> had shown that no process whose ultimate outcome is the capture of a particle by a Kerr black hole can result in the decrease of a certain quantity which he named the irreducible mass of the black hole,  $M_{ir}$ . In fact, most processes result in an increase in  $M_{ir}$  with the exception of a very special class of limiting processes, called reversible processes, which leave  $M_{ir}$  unchanged. It turns out that  $M_{ir}$  is proportional to the square root of the black hole's area<sup>5,6</sup> [see (1)]. Thus Christodoulou's result implies that the area increases in most processes, and thus it supports the conjecture of Floyd and Penrose. Christodoulou's conclusion is also valid for charged Kerr black holes.<sup>4,6</sup>

By an approach radically different from Christodoulou's, Hawking<sup>7</sup> has given a general proof that the black-hole surface area cannot decrease in any process. For a system of several black holes Hawking's theorem implies that the area of each individual black hole cannot decrease, and more-

over that when two black holes merge, the area of the resulting black hole (provided, of course, that one forms) cannot be smaller than the sum of initial areas.

It is clear that changes of a black hole generally take place in the direction of increasing area. This is reminiscent of the second law of thermodynamics which states that changes of a closed thermodynamic system take place in the direction of increasing entropy. The above comparison suggests that it might be useful to consider black-hole physics from a thermodynamic viewpoint: We already have the concept of energy in black-hole physics, and the above observation suggests that something like entropy may also play a role in it. Thus, one can hope to develop a thermodynamics for black holes – at least a rudimentary one. In this paper we show that it is possible to give a precise definition of black-hole entropy. Based on it we construct some elements of a thermodynamics for black holes.

There are some precedents to our considerations. The idea of making use of thermodynamic methods in black-hole physics appears to have been first considered by Greif.<sup>8</sup> He examined the possibility of defining the entropy of a black hole, but lacking many of the recent results in black-hole physics, he did not make a concrete proposal. More recently, Carter<sup>9</sup> has rederived the result of Christodoulou<sup>4,5</sup> that the irreducible mass of a Kerr black hole is unchanged in a reversible trans-

formation by applying to the black hole the criterion for a thermodynamically reversible transformation of a rigidly rotating star.<sup>10</sup> Carter's example shows the possibilities inherent in the use of thermodynamic arguments in black-hole physics.

In this paper we attempt a unification of black-hole physics with thermodynamics. In Sec. II we point out a number of analogies between black-hole physics and thermodynamics, all of which bring out the parallelism between black-hole area and entropy. In Sec. III, after a short review of elements of the theory of information, we discuss some features of black-hole physics from the point of view of information theory. We take the area of a black hole as a measure of its entropy—entropy in the sense of inaccessibility of information about its internal configuration. We go further in Sec. IV and propose a specific expression for black-hole entropy in terms of black-hole area. Earlier<sup>11,12</sup> we had proposed this same expression on different grounds; here we find the value of a previously unknown constant by means of an argument based on information theory. In Sec. V we propose a generalization of the second law of thermodynamics applicable to black-hole physics: When some common entropy goes down a black hole, *the black-hole entropy plus the common entropy in the black-hole exterior never decreases.*<sup>11,12</sup>

In Secs. VI and VII we construct several examples which provide support for the generalized second law. In addition, we analyze in Sec. VII a thought experiment proposed by Geroch<sup>13</sup> in which, with the help of a black hole, heat is apparently converted entirely into work in violation of the second law. We show that, in fact, due to fundamental physical limitations the conversion efficiency is somewhat smaller than unity. Moreover, the efficiency is no greater than the maximum efficiency allowed by thermodynamics for the heat engine which is equivalent to the Geroch process, so that this process cannot be regarded as violating the second law.

## II. ANALOGIES BETWEEN BLACK-HOLE PHYSICS AND THERMODYNAMICS

We have already mentioned the resemblance between the tendency of black-hole area to increase, and the tendency of entropy to increase. Changes of a black hole or of a system of black holes select a preferred direction in time: that in which the black-hole area increases. Likewise, changes of a closed thermodynamic system select a preferred direction in time: that in which the entropy increases. This parallelism between black-hole area and entropy goes even deeper.

Black-hole area turns out to be as intimately related to the degradation of energy as is entropy. In thermodynamics the statement "the entropy has increased" implies that a certain quantity of energy has been degraded, i.e., that it can no longer be transformed into work. Now, as Christodoulou has emphasized,<sup>4,5</sup> the irreducible mass  $M_{ir}$  of a Kerr black hole, which is related to the surface area  $A$  of the black hole by<sup>14</sup>

$$M_{ir} = (A/16\pi)^{1/2}, \quad (1)$$

represents energy which cannot be extracted by means of Penrose processes.<sup>3</sup> In this sense it is inert energy which cannot be transformed into work. Thus, an increase in  $A$ , and hence in  $M_{ir}$ , corresponds to the degradation (in the thermodynamic sense) of some of the energy of the black hole.

The irreducible mass of a Schwarzschild black hole is just equal to its total mass. Thus, no energy can be extracted from such a black hole by means of Penrose processes. However, the merger of two Schwarzschild black holes can yield energy in the form of gravitational waves.<sup>7</sup> The only restriction on the process is that the total black-hole area must not decrease as a result of the merger.<sup>7</sup> However, the sum of the irreducible masses of individual black holes may (in fact, does) decrease. We see that for a system of several black holes the degraded energy  $E_d$  is more appropriately given by

$$E_d = (\sum A/16\pi)^{1/2} = (\sum M_{ir}^2)^{1/2} \quad (2)$$

than by  $\sum M_{ir}$ . According to this formula the degraded energy of a system of black holes is smaller than the sum of degraded energies of the black holes considered separately. Thus by combining Schwarzschild black holes which are already "dead," one can still obtain energy.<sup>7</sup> Analogously, by allowing two thermodynamic systems which are separately in equilibrium to interact, one can obtain work, whereas each system by itself could have done no work. From the above observations the parallelism between black-hole area and entropy is again evident.

We shall now construct the black-hole analog of the thermodynamic expression

$$dE = TdS - PdV. \quad (3)$$

For convenience we shall from now on write all our equations in terms of the "rationalized area" of a black hole  $\alpha$  defined by

$$\alpha = A/4\pi. \quad (4)$$

Consider a Kerr black hole of mass  $M$ , charge  $Q$ , and angular momentum  $\vec{L}$ . (3-vectors here refer to components with respect to the Euclidean frame at infinity.) Its rationalized area is given by<sup>5,7</sup>

$$\begin{aligned}\alpha &= r_+^2 + a^2 \\ &= 2Mr_+ - Q^2,\end{aligned}\quad (5)$$

where

$$\vec{a} = \vec{L}/M,\quad (6)$$

$$r_+ = M \pm (M^2 - Q^2 - a^2)^{1/2}.\quad (7)$$

Differentiating (5) and solving for  $dM$  we obtain

$$dM = \Theta d\alpha + \vec{\Omega} \cdot d\vec{L} + \Phi dQ,\quad (8)$$

where

$$\Theta \equiv \frac{1}{4} (r_+ - r_-)/\alpha,\quad (9a)$$

$$\vec{\Omega} \equiv \vec{a}/\alpha,\quad (9b)$$

$$\Phi \equiv Qr_+/\alpha.\quad (9c)$$

In (8) we have the black-hole analog of the thermodynamic expression (3): The terms  $\vec{\Omega} \cdot d\vec{L}$  and  $\Phi dQ$  clearly represent the work done on the black hole by an external agent who increases the black hole's angular momentum and charge by  $d\vec{L}$  and  $dQ$ , respectively. Thus  $\vec{\Omega} \cdot d\vec{L} + \Phi dQ$  is the analog of  $-PdV$ , the work done on a thermodynamic system. Comparing our expression for work with the expressions for work done on rotating<sup>15</sup> and charged<sup>16</sup> bodies, we see that  $\vec{\Omega}$  and  $\Phi$  play the roles of rotational angular frequency and electric potential of the black hole, respectively.<sup>5,9</sup> The  $\alpha$  in (8) resembles the entropy  $S$  in (3) as we have noted before: For any change of the black hole  $d\alpha \geq 0$ ,<sup>5,7</sup> while for any change of a closed thermodynamic system  $dS \geq 0$ . Moreover, it is clear from (7) and (9a) that  $\Theta$ , the black-hole analog of temperature  $T$ , is non-negative just as  $T$  is. From the above observations the formal correspondence between (3) and (8) is evident.

All the analogies we have mentioned are suggestive of a connection between thermodynamics and black-hole physics in general, and between entropy and black-hole area in particular. But so far the analogies have been of a purely formal nature, primarily because entropy and area have different dimensions. We shall remedy this deficiency in Sec. IV by constructing out of black-hole area an expression for black-hole entropy with the correct dimensions. Preparatory to this we shall now look

at black-hole physics from the point of view of the theory of information.

### III. INFORMATION AND BLACK-HOLE ENTROPY

The connection between entropy and information is well known.<sup>17,18</sup> The entropy of a system measures one's uncertainty or lack of information about the actual internal configuration of the system. Suppose that all that is known about the internal configuration of a system is that it may be found in any of a number of states with probability  $p_n$  for the  $n$ th state. Then the entropy associated with the system is given by Shannon's formula<sup>17,18</sup>

$$S = - \sum_n p_n \ln p_n .\quad (10)$$

This formula is uniquely determined by a few very general requirements which are imposed in order that  $S$  have the properties expected of a measure of uncertainty.<sup>17</sup>

It should be noticed that the above entropy is dimensionless. This simply means that we choose to measure temperature in units of energy. Boltzmann's constant is then dimensionless.

Whenever new information about the system becomes available, it may be regarded as imposing some constraints on the probabilities  $p_n$ . For example, the information may be that several of the  $p_n$  are, in fact, zero. Such constraints on the  $p_n$  always result in a decrease in the entropy function.<sup>18</sup> This property is formalized by the relation<sup>17,18</sup>

$$\Delta I = -\Delta S,\quad (11)$$

where  $\Delta I$  is the new information which corresponds to a decrease  $\Delta S$  in one's uncertainty about the internal state of the system. Equation (11) is the basis for Brillouin's identification of information with negative entropy.<sup>18</sup> All the above comments apply to such diverse systems as a quantity of gas in a box or a telegram. A familiar example of the relation between a gain of information and a decrease in entropy is the following. Some ideal gas in a container is compressed isothermally. It is well known that its entropy decreases. On the other hand, one's information about the internal configuration of the gas increases: After the compression the molecules of the gas are more localized, so that their positions are known with more accuracy than before the compression.

The second law of thermodynamics is easily understood in the context of information theory. The entropy of a thermodynamic system which is not in equilibrium increases because information

about the internal configuration of the system is being lost during its evolution as a result of the washing out of the effects of the initial conditions. It is possible for an exterior agent to cause a decrease in the entropy of a system by first acquiring information about the internal configuration of the system. The classic example of this is that of Maxwell's demon.<sup>18</sup> But information is never free. In acquiring information  $\Delta I$  about the system, the agent inevitably causes an increase in the entropy of the rest of the universe which exceeds  $\Delta I$ .<sup>18</sup> Thus, even though the entropy of the system decreases in accordance with (11), the over-all entropy of the universe increases in the process.

The conventional unit of information is the "bit" which may be defined as the information available when the answer to a yes-or-no question is precisely known (zero entropy). According to the scheme (11) a bit is also numerically equal to the maximum entropy that can be associated with a yes-or-no question, i.e., the entropy when no information whatsoever is available about the answer. One easily finds that the entropy function (10) is maximized when  $p_{yes} = p_{no} = \frac{1}{2}$ . Thus, in our units, one bit is equal to  $\ln 2$  of information.

Let us now return to our original subject, black holes. In the context of information a black hole is very much like a thermodynamic system. The entropy of a thermodynamic system in equilibrium measures the uncertainty as to which of all its internal configurations compatible with its macroscopic thermodynamic parameters (temperature, pressure, etc.) is actually realized. Now, just as a thermodynamic system in equilibrium can be completely described macroscopically by a few thermodynamic parameters, so a bare black hole in equilibrium (Kerr black hole) can be completely described (insofar as an exterior observer is concerned) by just three parameters: mass, charge, and angular momentum.<sup>1</sup> Black holes in equilibrium having the same set of three parameters may still have different "internal configurations." For example, one black hole may have been formed by the collapse of a normal star, a second by the collapse of a neutron star, a third by the collapse of a geon. These various alternatives may be regarded as different possible internal configurations of one and the same black hole characterized by their (common) mass, charge, and angular momentum. It is then natural to introduce the concept of black-hole entropy as the measure of the *inaccessibility* of information (to an exterior observer) as to which particular internal configuration of the black hole is actually realized in a given case.

At the outset it should be clear that the black-hole entropy we are speaking of is *not* the thermal

entropy inside the black hole. In fact, our black-hole entropy refers to the equivalence class of all black holes which have the same mass, charge, and angular momentum, not to one particular black hole. What are we to take as a measure of this black-hole entropy? The discussion of Sec. II predisposes us to single out black-hole area. But to be more general we shall only assume that the entropy of a black hole,  $S_{bh}$ , is some *monotonically increasing* function of its rationalized area:

$$S_{bh} = f(\alpha). \quad (12)$$

Although our motivating discussion for the introduction of the concept of the black-hole entropy made use of the specific properties of stationary black holes, we shall take (12) to be valid for any black hole, including a dynamically evolving one, since the surface area is well defined for any black hole. This choice is supported by the following observations.

As mentioned earlier, the entropy of an evolving thermodynamic system increases due to the gradual loss of information which is a consequence of the washing out of the effects of the initial conditions. Now, as a black hole approaches equilibrium, the effects of the initial conditions are also washed out (the black hole loses its hair)<sup>1</sup>; only mass, charge, and angular momentum are left as determinants of the black hole at late times. We would thus expect that the loss of information about initial peculiarities of the hole will be reflected in a gradual increase in  $S_{bh}$ . And indeed Eq. (12) predicts just this; by Hawking's theorem  $S_{bh}$  increases monotonically as the black hole evolves. This agreement is evidence in favor of the choice (12).

We mentioned earlier that the possibility of causing a decrease in the entropy of a thermodynamic system goes hand in hand with the possibility of obtaining information about its internal configuration. By contrast, an exterior agent cannot acquire any information about the interior configuration of a black hole. The one-way membrane nature of the event horizon prevents him from doing so.<sup>1</sup> Therefore, we do not expect an exterior agent to be able to cause a decrease in the black hole's entropy. Equation (12) is in agreement with this expectation; by Hawking's theorem  $S_{bh}$  never decreases. Here we have a new piece of evidence in favor of the choice (12).

One possible choice for  $f$  in (12),  $f(\alpha) \propto \alpha^{1/2}$ , is untenable on various grounds. Consider two black holes which start off very distant from each other. Since they interact weakly we can take the total black-hole entropy to be the sum of the  $S_{bh}$  of each black hole. The black holes now fall together,

merge, and form a black hole which settles down to equilibrium. In the process no information about the black-hole interior can become available; on the contrary, much information is lost as the final black hole "loses its hair." Thus, we expect the final black-hole entropy to exceed the initial one. By our assumption that  $f(\alpha) \propto \alpha^{1/2}$ , this implies that the irreducible mass [see (1)] of the final black hole exceeds the sum of irreducible masses of the initial black holes. Now suppose that all three black holes are Schwarzschild ( $M = M_{\text{ir}}$ ). We are then confronted with the prediction that the final black-hole mass exceeds the initial one. But this is nonsense since the total black-hole mass can only decrease due to gravitational radiation losses. We thus see that the choice  $f(\alpha) \propto \alpha^{1/2}$  is untenable.

The next simplest choice for  $f$  is

$$f(\alpha) = \gamma \alpha, \quad (13)$$

where  $\gamma$  is a constant. Repetition of the above argument for this new  $f$  leads to the conclusion that the final black-hole area must exceed the total initial black-hole area. But we know this to be true from Hawking's theorem.<sup>7</sup> Thus the choice (13) leads to no contradiction. Therefore, we shall adopt (13) for the moment; later on we shall exhibit some more positive evidence in its favor.

Comparison of (12) and (13) shows that  $\gamma$  must have the units of  $(\text{length})^{-2}$ . But there is no constant with such units in classical general relativity. If in desperation we appeal to quantum physics we find only one truly universal constant with the correct units<sup>14</sup>:  $\hbar^{-1}$ , that is, the reciprocal of the Planck length squared. (Compton wavelengths are not universal, but peculiar to this or that particle; they clearly have no bearing on the problem.) We are thus compelled to write (12) as

$$S_{\text{bh}} = \eta \hbar^{-1} \alpha, \quad (14)$$

where  $\eta$  is a dimensionless constant which we may expect to be of order unity. This expression was first proposed by us earlier<sup>11,12</sup> from a different point of view.

We need not be alarmed at the appearance of  $\hbar$  in the expression for black-hole entropy. It is well known<sup>15</sup> that  $\hbar$  also appears in the formulas for entropy of many thermodynamic systems that are conventionally regarded as classical, for example, the Boltzmann ideal gas. This is a reflection of the fact that entropy is, in a sense, a count of states of the system, and the underlying states of any system are always quantum in nature. It is thus not totally unexpected that  $\hbar$  appears in (14). These observations also suggest that it would be

somewhat pretentious to attempt to calculate the precise value of the constant  $\eta \hbar^{-1}$  without a full understanding of the quantum reality which underlies a "classical" black hole. Since there is no hope at present of obtaining such an understanding, we bypass the issue, and in the next section we use a semiclassical argument to arrive at a value for  $\eta \hbar^{-1}$  which should be quite close to the correct one.

#### IV. EXPRESSION FOR BLACK-HOLE ENTROPY

In our attempt to obtain a value for  $\eta \hbar^{-1}$  we shall employ the following argument. We imagine that a particle goes down a Kerr black hole. As it disappears some information is lost with it. According to the discussion of Sec. III we expect the black-hole entropy, as the measure of inaccessible information, to reflect the loss of the information associated with the particle by increasing by an appropriate amount. How much information is lost together with the particle? The amount clearly depends on how much is known about the internal state of the particle, on the precise way in which the particle falls in, etc. But we can be sure that the absolute minimum of information lost is that contained in the answer to the question "does the particle exist or not?" To start with, the answer is known to be yes. But after the particle falls in, one has no information whatever about the answer. This is because from the point of view of this paper, one knows nothing about the physical conditions inside the black hole, and thus one cannot assess the likelihood of the particle continuing to exist or being destroyed. One must, therefore, admit to the loss of one bit of information (see Sec. III) at the very least.

Our plan, therefore, is to compute the minimum possible increase in the black hole's area which results from the disappearance of a particle down the black hole, then to compute the corresponding minimum possible increase of black-hole entropy by means of our original formula (12), and finally to identify this increase in entropy with the loss of one bit of information in accordance with the scheme (11). If our procedure is reasonable we should then recover the functional form of  $f$  given by (13), together with a definite value for  $\gamma$ .

There are many ways in which a particle can go down a black hole, all leading to varying increases in black-hole area. We are interested in that method for inserting the particle which results in the smallest increase. This method has already been discussed by Christodoulou<sup>4-6</sup> in connection with his introduction of the concept of irreducible mass. The essence of Christodoulou's method is that if a freely falling point particle is captured by

a Kerr black hole from a turning point in its orbit, then the irreducible mass and, consequently, the area of the hole are left unchanged. For reasons that will become clear presently we wish to allow the particle to have a nonzero radius. As shown in Appendix A, Christodoulou's method can be generalized easily so as to allow for this, as well as for the possibility that the particle is brought to the horizon by some method other than by free fall. We find in Appendix A that the increase in area for the generalized Christodoulou process is no longer precisely zero. But interestingly enough, the minimum increase in rationalized area,  $(\Delta\alpha)_{\min}$ , turns out to be independent of the parameters of the black hole. For a spherical particle of rest mass  $\mu$ , and proper radius  $b$ ,

$$(\Delta\alpha)_{\min} = 2\mu b. \quad (15)$$

For a point particle  $(\Delta\alpha)_{\min} = 0$ ; this is Christodoulou's result.

Expression (15) gives the minimum possible increase in black-hole area that results if a given particle is added to a Kerr black hole. We can try to make  $(\Delta\alpha)_{\min}$  smaller by making  $b$  smaller. However, we must remember that  $b$  can be no smaller than the particle's Compton wavelength  $\hbar\mu^{-1}$ , or than its gravitational radius  $2\mu$ , whichever is the larger. The Compton wavelength is the larger for  $\mu < 2^{-1/2}\hbar^{1/2}$ , and the gravitational radius is the larger for  $\mu > 2^{-1/2}\hbar^{1/2}(2^{-1/2}\hbar^{1/2} \approx 10^{-5} g)$ . Thus, if  $\mu < 2^{-1/2}\hbar^{1/2}$ , then  $2\mu b$  can be as small as  $2\mu\hbar\mu^{-1} = 2\hbar$ . But if  $\mu > 2^{-1/2}\hbar^{1/2}$ , then  $2\mu b$  can be no smaller than  $4\mu^2 > 2\hbar$ . We conclude that quantum effects set a lower bound of  $2\hbar$  on the increase of the rationalized area of a Kerr black hole when it captures a particle. Moreover, this limit can be reached only for a particle whose dimension is given by its Compton wavelength. Of course, only such an "elementary particle" can be regarded as having no internal structure. Therefore, the loss of information associated with the loss of such a particle should be minimum. And indeed we find that the increase in black-hole entropy is smallest for just such a particle. This supports our view that  $2\hbar$  is the increase in rationalized area associated with the loss of one bit of information.

Following our program we shall equate the minimum increase in black-hole entropy,  $(\Delta S_{bh})_{\min} = 2\hbar df/d\alpha$ , with  $\ln 2$ , the entropy increase associated with the loss of one bit of information. Integration of the resulting equation gives  $f(\alpha) = (\frac{1}{2}\ln 2)\hbar^{-1}\alpha$ . Thus, we have arrived again at (13) by an alternate route, and have obtained the value of  $\gamma$  into the bargain. We now have

$$S_{bh} = (\frac{1}{2}\ln 2)\hbar^{-1}\alpha, \quad (16)$$

which is of the same form as (14). Our argument has determined the dependence of  $S_{bh}$  on  $\alpha$  in a straightforward manner. However, our value  $\eta = \frac{1}{2}\ln 2$  might presumably be challenged on the grounds that it follows from the assumption that the smallest possible radius of a particle is precisely equal to its Compton wavelength whereas the actual radius is not so sharply defined. Nevertheless, it should be clear that if  $\eta$  is not exactly  $\frac{1}{2}\ln 2$ , then it must be very close to this, probably within a factor of two. This slight uncertainty in the value of  $\eta$  is the price we pay for not giving our problem a full quantum treatment. However, in what follows we shall suppose that  $\eta = \frac{1}{2}\ln 2$ . Examples to be given later will show that this value leads to no contradictions.

How is the entropy of a system of several black holes defined? It is natural to define it as the sum of individual black-hole entropies. Then Hawking's theorem tells us that the total black-hole entropy of the system cannot decrease. But this is just what we would expect since the information lost down the black holes is unrecoverable. This observation lends support to our choice.

In conventional units (16) takes the form

$$\begin{aligned} S_{bh} &= (\frac{1}{2}\ln 2/4\pi)kc^3\hbar^{-1}G^{-1}A \\ &= (1.46 \times 10^{48} \text{ erg } ^{\circ}\text{K}^{-1} \text{ cm}^{-2})A, \end{aligned} \quad (17)$$

where  $k$  is Boltzmann's constant. We see that the entropy of a black hole is enormous. For example, a black hole of one solar mass would have  $S_{bh} \approx 10^{60} \text{ erg } ^{\circ}\text{K}^{-1}$ . By comparison the entropy of the sun is  $S \approx 10^{42} \text{ erg } ^{\circ}\text{K}^{-1}$ ; those of a white dwarf or a neutron star of one solar mass even smaller. The large numerical value of black-hole entropy serves to dramatize the highly irreversible character of the process of black-hole formation. We may define a characteristic temperature for a Kerr black hole by the relation  $T_{bh}^{-1} = (\partial S_{bh}/\partial M)_L|_Q$  which is the analog of the thermodynamic relation  $T^{-1} = (\partial S/\partial E)_V$ . By using (8) and (16) we find

$$\begin{aligned} T_{bh} &= 2\hbar(\ln 2)^{-1}\Theta \\ &= (0.165 \text{ } ^{\circ}\text{K cm})(r_+ - r_-)(r_+^2 + a^2)^{-1}, \end{aligned} \quad (18)$$

where  $r_+$  and  $a$  are to be given in centimeters. We introduce this  $T_{bh}$  in anticipation of our discussion of an example in Sec. VII. But we emphasize that one should not regard  $T_{bh}$  as the temperature of the black hole; such an identification can easily lead to all sorts of paradoxes, and is thus not useful.

### V. THE GENERALIZED SECOND LAW

Suppose that a body containing some common entropy goes down a black hole. The entropy of the visible universe decreases in the process. It would seem that the second law of thermodynamics is transcended here in the sense that an exterior observer can never verify by direct measurement that the total entropy of the whole universe does not decrease in the process.<sup>19</sup> However, we know that the black-hole area "compensates" for the disappearance of the body by increasing irreversibly. It is thus natural to conjecture that the second law is not really transcended provided that it is expressed in a generalized form: *The common entropy in the black-hole exterior plus the black-hole entropy never decreases.* This statement means that we must regard black-hole entropy as a genuine contribution to the entropy content of the universe.

Support for the above version of the second law comes from the following argument. Suppose that a body carrying entropy  $S$  goes down a black hole (which may have existed previously or may be formed by the collapse of the body). The  $S$  is the uncertainty in one's knowledge of the internal configuration of the body. So long as the body was still outside the black hole, one had the option of removing this uncertainty by carrying out measurements and obtaining information up to the amount  $S$ . But once the body has fallen in, this option is lost; the information about the internal configuration of the body becomes truly inaccessible. We thus expect the black-hole entropy, as the measure of inaccessible information, to increase by an amount  $S$ . Actually, the increase in  $S_{bh}$  may be even larger because any information that was available about the body to start with will also be lost down the black hole. Therefore, if we denote by  $\Delta S_c$  the change in common entropy in the black-hole exterior ( $\Delta S_c \equiv -S$ ), then we expect that

$$\Delta S_{bh} + \Delta S_c = \Delta(S_{bh} + S_c) > 0. \quad (19)$$

This is just the generalized second law which we proposed above: The generalized entropy  $S_{bh} + S_c$  never decreases. Examples supporting this law will be given in Sec. VI-VII.

This is a good place to mention the question of fluctuations. We know that the common entropy of a closed thermodynamic system can decrease spontaneously as a result of statistical fluctuations, i.e., the second law, being a statistical law, is meaningful only if statistical fluctuations are small. Is black-hole entropy also subject to decreases of a statistical nature? Not classically — Hawking's theorem guarantees that. Quantum mechanically

there are two ways by which the black-hole entropy can undergo statistical decreases. One of them depends on the quantum fluctuations of the metric of the black hole which one has reasons to expect.<sup>1</sup> Such fluctuations would be reflected in small random fluctuations in the area, and thus in the entropy of the black hole, and some of these fluctuations would be expected to be decreases in entropy. However, even if one regards a black hole as a purely classical object, it is still possible for its area and entropy to undergo small decreases when the black hole absorbs a single quantum under certain conditions.<sup>20</sup> However, the probability of such an event occurring in any given trial is very small. Therefore, the decrease in area and entropy is of a statistical nature, and is quite analogous to the decrease in entropy of a thermodynamic system due to statistical fluctuations. This discussion serves us warning that the law (19) is expected to hold only insofar as statistical fluctuations are negligible.

We noticed earlier (Sec. IV) the very large magnitude of black-hole entropy. In fact, one can say that the black-hole state is the maximum entropy state of a given amount of matter. The point is that in the gravitational collapse of a body into a black hole, the loss of information down the black hole is the maximum allowed by the laws of physics. Thus if the body collapses to form a Kerr black hole, all information about it is lost with the exception of mass, charge, and angular momentum.<sup>1</sup> These quantities are given in terms of Gaussian integrals,<sup>1</sup> and so information about them cannot be lost. But all other information about the body is eventually lost. Therefore, the resulting black hole must correspond to the maximum (generalized) entropy which can be associated with the given body.

### VI. EXAMPLES OF THE GENERALIZED SECOND LAW AT WORK

In the examples which follow we endeavor to subject the generalized second law to the most stringent test possible in each case by maximizing the entropy going down the black hole with a given body while minimizing the associated increase in black-hole entropy.

#### A. Harmonic Oscillator

As a first example we take an harmonic oscillator composed of two particles of rest mass  $\frac{1}{2}m$  each connected by a nearly massless spring of spring constant  $K$ . We imagine the oscillator to be enclosed in a spherical box and to be maintained at temperature  $T$ . We assume for simplicity that conditions are such that the oscillator

is nonrelativistic ( $T \ll m$ ). Let  $\omega$  be the vibrational frequency of the oscillator. Then the (normalized) probability that the oscillator is in its  $n$ th quantum state is given by the canonical distribution

$$p_n = (1 - e^{-x})^{-1} e^{-nx}, \quad x = \hbar\omega/T. \quad (20)$$

The entropy of the oscillator as computed from (10) is

$$S = x(e^x - 1)^{-1} - \ln(1 - e^{-x}), \quad (21)$$

and the mean vibrational energy,

$$\langle E \rangle \equiv \sum p_n (n + \frac{1}{2}) \hbar\omega,$$

is

$$\langle E \rangle = [(e^x - 1)^{-1} + \frac{1}{2}] \hbar\omega. \quad (22)$$

We remark that the thermal distribution (20) maximizes the entropy of the oscillator for given  $\langle E \rangle$ , and is thus ideally suited to our plan for subjecting the generalized second law to the most stringent test possible.

Suppose that the box goes down a Kerr black hole. The corresponding increase in black-hole entropy cannot be smaller than the lowest limit derived by the method of Appendix A. From (15) and (16) we have  $\Delta S_{bh} \geq \mu b \hbar^{-1} \ln 2$ , where  $b$  is the outer radius of the box and  $\mu$  is its total rest mass. Clearly  $b$  must be at least as large as half of the mean value  $\langle y \rangle$  of the separation of the two masses  $y$ . And  $\langle y \rangle$  in turn must clearly be larger than  $\Delta y$ , the root mean square of the thermal oscillation of  $y$  [ $(\Delta y)^2 = \langle (y - \langle y \rangle)^2 \rangle$ ], so that  $y$  will always be positive. Now according to the (quantum) virial theorem  $\frac{1}{2}\langle E \rangle$  is equal to the mean potential energy of the oscillator  $\frac{1}{2}K(\Delta y)^2$ . Since the reduced mass of the oscillator is  $\frac{1}{4}m$ , we have  $K = \frac{1}{4}m\omega^2$ . We thus find from all the above that

$$b > \langle E \rangle^{1/2} m^{-1/2} \omega^{-1}.$$

Remembering that  $\mu > m + \langle E \rangle$  (because the box itself must have some mass) we obtain

$$\Delta S_{bh} > \langle E \rangle^{1/2} m^{-1/2} (\hbar\omega)^{-1} (m + \langle E \rangle) \ln 2. \quad (23)$$

We assume that the entropy given by (21) is the only contribution to the entropy in the box. This amounts to neglecting the contribution of the black body radiation in the box, etc., a sensible procedure if  $T$  is not very high. Then  $\Delta S_c = -S$  and we have

$$\begin{aligned} \Delta(S_{bh} + S_c) &> \xi^{-1/2} (1 + \xi) [\frac{1}{2} + (e^x - 1)^{-1}] \ln 2 \\ &\quad - x(e^x - 1)^{-1} + \ln(1 - e^{-x}), \end{aligned} \quad (24)$$

where we have introduced the notation  $\xi = m\langle E \rangle^{-1}$  and used Eqs. (21) and (22) for  $S$  and  $\langle E \rangle$ . We now show that  $\Delta(S_{bh} + S_c) > 0$  as required by the generalized second law. The expression in (24) regarded as a function of  $x$  for given  $\xi$  has a single minimum at

$$\begin{aligned} x &= x_m \\ &\equiv \xi^{-1/2} (1 + \xi) \ln 2 \\ &\equiv \frac{1}{2} x_m + \ln[1 - \exp(-x_m)]. \end{aligned}$$

which has the value

Our assumption that the oscillator is nonrelativistic means that  $\xi \gg 1$ , and hence that  $x_m \gg 2 \ln 2$ . Under these conditions the minimum is positive (in fact, it is positive for  $\xi \geq 1$ ). It follows immediately that  $\Delta(S_{bh} + S_c)$  is positive for all  $x$  and all  $\xi$  which are compatible with the requirement of a nonrelativistic oscillator. The generalized second law is obeyed over the entire regime for which our treatment is valid.

### B. Beam of Light

As a second example we consider a beam of light which is aimed at a Kerr black hole. This example is particularly interesting because it shall bring us face to face with the issue of fluctuations as a limitation on the applicability of the second law. We shall restrict our attention only to those cases for which geometrical optics is a valid approximation. We shall thus represent the path of the beam by a null geodesic in the Kerr background.

We shall take it that the beam is thermalized at a certain temperature  $T$ . This implies that its entropy is a maximum for given energy. The entropy is easily calculated; in fact, the entropy and energy for each mode in the beam are given by the same expressions (21) and (22) which apply to a harmonic oscillator, except that one must omit the zero-point energy term  $\frac{1}{2}\hbar\omega$ . The total entropy  $S$  and mean energy  $\langle E \rangle$  of the beam are obtained by integrating these expressions weighed by the conventional density of states

$$\rho = 2\omega^2 (2\pi)^{-3} V d\Omega \quad (25)$$

over all  $\omega$ . In (25)  $V$  is the volume of the beam and  $d\Omega$  is the solid angle it subtends. Integrating by parts the expression for  $S$ , one easily obtains

the relation

$$S = \frac{4}{3} \langle E \rangle T^{-1}, \quad (26)$$

which, not surprisingly, is identical to that for radiation inside a black-body cavity of temperature  $T$ .<sup>15</sup> [In Ref. 12 (26) was given with an incorrect numerical factor.]

As the beam nears the black hole, it is deflected by the gravitational field. Insofar as its effects on electromagnetic radiation are concerned, a stationary gravitational field can be mocked up by an appropriate nonabsorbing refractive medium in flat spacetime.<sup>21</sup> But the propagation of a beam of light through such a medium is a reversible process.<sup>22</sup> We infer from this that the entropy of the beam will remain unchanged as the beam nears the black hole. Thus the entropy change of the visible universe when the beam goes down the hole is just

$$\Delta S_c = -\frac{4}{3} \langle E \rangle T^{-1}. \quad (27)$$

What is the increase in black-hole entropy associated with the process? From (8) we see that the increase in  $\alpha$  is minimized when the angular momentum that the hole gains from the beam is maximized for given  $\langle E \rangle$ . Now, the gain in angular momentum is limited because the beam will not be captured if it carries too much angular momentum. In Appendix B we take this into account in calculating (in the geometrical optics limit) the minimum possible increase in  $\alpha$  for given  $\langle E \rangle$  of the beam. We find that  $\Delta\alpha \geq \beta M \langle E \rangle$  where  $\beta$  ranges from 8 for the case of a Schwarzschild hole to  $4(1-\sqrt{3}/2)$  for the case of an extreme Kerr hole, this last value being the smallest possible  $\beta$ . From (16) it follows that

$$\Delta S_{bh} \geq (\frac{1}{2}\beta \ln 2) M \hbar^{-1} \langle E \rangle. \quad (28)$$

Our assumption (Appendix B) that geometrical optics is always applicable means that the bulk of wavelengths in the beam are much shorter than the characteristic dimension of the hole  $\approx M$ . Thus, if  $\omega_c$  is some characteristic frequency in the beam, then we require that  $\omega_c \gg M^{-1}$ . From the form of the Planck spectrum (22) we see that  $\hbar\omega_c \approx T$ ; therefore (27) tells us that

$$|\Delta S_c| \ll \frac{4}{3} M \hbar^{-1} \langle E \rangle. \quad (29)$$

Comparison of (28) and (29) shows that a violation of the generalized second law (19) cannot arise in the regime under consideration.

In the above discussion the condition that geometrical optics be applicable prevented us from taking  $T$  to be arbitrarily small. As a result it

turned out to be impossible for  $|\Delta S_c|$  to exceed  $\Delta S_{bh}$ , and so a violation of the second law was ruled out. But there is a way to circumvent the restriction on  $T$ . One simply selects the temperature  $T$  (arbitrarily) to be as small as one pleases, and arranges for all frequencies  $\omega < \omega'_c$  to be filtered out of the beam. Here  $\omega'_c \gg M^{-1}$  is a definite frequency unrelated to  $T$ . It should be clear that geometrical optics will be a valid approximation for this case also, so that we may take over the result (28). But the result (27) must be modified since we are here dealing with a truncated frequency spectrum. We are mostly interested in the regime  $T \ll \hbar \omega'_c$ . Then for all frequencies in the beam  $x = \hbar\omega/T \gg 1$ . It follows from (21) and (22) that for each mode the entropy to energy ratio is  $T^{-1}$  ( $S \approx xe^{-x}$ ,  $\langle E \rangle \approx \hbar\omega e^{-x}$ ). Therefore instead of (27) we have

$$\Delta S_c = -\langle E \rangle T^{-1}. \quad (30)$$

It now appears that if

$$T < T_c \equiv \hbar(\frac{1}{2}\beta M \ln 2)^{-1},$$

then  $\Delta S_{bh} + \Delta S_c$  will be negative in contradiction with the generalized second law.

The resolution of the above paradox is that in the regime  $T < T_c$  statistical fluctuations are already dominant so that our entire picture of the process is invalid. To verify the importance of fluctuations we calculate the mean number of quanta  $N$  in the beam by integrating the mean number of quanta per mode,  $(e^x - 1)^{-1}$ , weighed by the density of states (25) over all  $\omega > \omega'_c$ . For  $T = T_c$  we get (recall that  $\hbar\omega'_c/T_c \gg 1$  by our assumptions)

$$N \approx \frac{V}{M^3} \frac{d\Omega}{4\pi} \delta^{-3} (\delta M \omega'_c)^2 \exp(-\delta M \omega'_c), \quad (31)$$

where  $\delta \equiv \frac{1}{2}\beta \ln 2$  ( $0.2 \leq \delta \leq 2.8$ ). It is clear that for any beam aimed at the black hole  $d\Omega/4\pi \ll 1$ . Recalling that  $M\omega'_c \gg 1$  by assumption, we see from (31) that each quantum occupies a mean volume much larger than  $M^3$ . But the cross section of the beam must be smaller than  $\sim M^2$  if the beam is to go down the black hole. Thus the mean separation between quanta is much larger than  $M$ , the characteristic dimension of the black hole. In case  $T < T_c$  the above effect is even more accentuated.

We conclude that in the regime  $T \leq T_c$  for which the second law (19) appears to break down, our description of the process as a continuous beam going down the black hole is invalidated by the large fluctuations in the concentration of energy in the beam (or equivalently, the large fluctuations in the energy of each section of the beam). In this

regime  $\langle E \rangle$  is no longer a good measure of the actual energy. It appears, therefore, that statistical fluctuations are responsible for the breakdown of the second law in the context in which we have applied it here. But we can demonstrate that the law has not lost all its meaning by adopting a point of view more suitable to the circumstances at hand than the one used above.

We take the point of view that quanta are going down the black hole one at a time, rather than in a continuous stream. Thus we must check the validity of the generalized second law for the infall of each quantum. The analysis of Appendix B still leads to formula (28) for the increase in black-hole entropy except that  $\langle E \rangle$  is replaced by  $\hbar\omega$ , the energy of the quantum. To compute the common entropy going down the black hole we reason as follows. From our point of view a quantum of definite frequency is going down the black hole. Thus we are no longer dealing with the probability distribution (20); instead we shall ascribe probability  $\frac{1}{2}$  to each of the two possible polarizations of the quantum. Then according to (10) the entropy associated with the quantum is  $\ln 2$ . Therefore,

$$\Delta S_{bh} + \Delta S_c \geq (\frac{1}{2}\beta \ln 2)M\omega - \ln 2. \quad (32)$$

Since  $\frac{1}{2}\beta \geq 0.268$ , and since we are assuming that  $M\omega \gg 1$ , we see that  $\Delta S_{bh} + \Delta S_c$  is in fact positive: The generalized second law is upheld for the infall of each quantum.

## VII. A PERPETUAL MOTION MACHINE USING A BLACK HOLE?

Geroch<sup>13</sup> has described a procedure using a black hole which appears to violate the second law of thermodynamics by converting heat into work with unit efficiency. He envisages a box filled with black-body radiation which is slowly lowered by means of a string from far away down to the horizon of a black hole, at which point its energy as measured from infinity vanishes. Therefore, if the box's rest mass is  $\mu$ , then the agent lowering the string obtains work equal to  $\mu$  out of the process. The box is then allowed to emit into the black-hole radiation of (proper) energy  $\Delta\mu$ . Finally, the agent retrieves the box; since its rest mass is now  $\mu - \Delta\mu$ , he must do work  $\mu - \Delta\mu$  to accomplish this. Therefore, in the whole process the agent obtains net work  $\Delta\mu$  at the expense of heat  $\Delta\mu$  — conversion with unit efficiency. We shall now show that, in fact, due to fundamental physical limitations, the efficiency of the Geroch process is slightly smaller than unity, so that no violation of the second law is entailed here.

The box under consideration must have a non-zero radius (see below). Because of this its energy as measured from infinity is never quite zero when it is as close to the horizon as it can possibly be. We shall assume that the box is in the shape of a sphere of radius  $b$ . Then according to the analysis of Appendix C the minimum value of the energy is

$$E = 2\mu b\Theta, \quad (33)$$

where  $\Theta$  is defined by (9a). It follows that in lowering the box from infinity to the horizon, the agent obtains only work  $\mu(1-2b\Theta)$  rather than  $\mu$ . After the box has radiated into the black hole, its rest mass becomes  $\mu - \Delta\mu$  and according to (33) its energy at the horizon is just  $2(\mu - \Delta\mu)b\Theta$ . Thus the agent must do work  $(\mu - \Delta\mu)(1-2b\Theta)$  to retrieve the box to infinity where its energy is  $\mu - \Delta\mu$ . Therefore, in the over-all process the agent obtains net work  $\Delta\mu(1-2b\Theta)$  in exchange for the expenditure of heat  $\Delta\mu$ . The efficiency of conversion is

$$\epsilon = 1 - 2b\Theta, \quad (34)$$

which is smaller than unity. In practical situations  $b \ll r_+$  so that  $b\Theta \ll 1$  and the efficiency can be quite near to unity. But the departure of  $\epsilon$  from unity, albeit small, serves to resolve the problem raised by Geroch's example: There is no violation of the Kelvin statement of the second law.<sup>23</sup>

We must now explain why  $b$  cannot be arbitrarily small. Physically the reason is that the box must be large enough for the wavelengths characteristic of radiation of some temperature  $T$  to fit into it. More formally we can argue as follows. The frequency of the photon ground state associated with the box,  $\omega_0$ , cannot exceed that frequency  $\omega_p$  at which the Planck photon-number spectrum

$$\propto \omega^2 [\exp(\hbar\omega/T) - 1]^{-1}$$

peaks. Otherwise the frequencies of all photon states would lie in the exponential tail of the spectrum, the occupation number of each state would be small, and the resulting large fluctuations would make the concept of temperature meaningless. We have the conventional relation  $\omega_0 b' = \pi$ , where  $b'$  is the interior radius of the box ( $b' < b$ ), and we easily find that  $\hbar\omega_p < 2T$ . Therefore,  $\omega_0 < \omega_p$  implies that  $b > \pi\hbar/2T$ . It is thus clear that there is a lower limit for  $b$ .

We may write the efficiency (34) in a more transparent form by recalling that  $\Theta = \frac{1}{2}T_{bh} \ln 2/\hbar$  [see (18)], where  $T_{bh}$  is the characteristic temperature associated with the black hole. Since  $b > \pi\hbar/2T$  we

find that

$$\epsilon < 1 - T_{bh}/T. \quad (35)$$

We now recall that the efficiency of a heat engine operating between two reservoirs, one at temperature  $T$  and the second at temperature  $T_{bh} < T$ , is restricted by  $\epsilon < 1 - T_{bh}/T$ . We thus see that the Geroch process is no more efficient than its "equivalent reversible heat engine." This observation makes it evident again that Geroch's process is not in violation of the second law of thermodynamics. Finally, we wish to remark that since our primary formula (33) is valid only when the box is small compared to the black hole ( $b\Theta \ll 1$ ), we can vouch for the validity of (35) only when  $T_{bh} \ll T$ . However, due to the smallness of  $T_{bh}$  this condition will be satisfied in all cases of practical interest.

We now verify that the Geroch process is in accord with the generalized second law (19). We mentioned earlier that the agent obtains work  $\Delta\mu(1-2b\Theta)$  for a decrease  $\Delta\mu$  in the rest mass of the box. This means that the black hole's mass must increase by  $2\Delta\mu b\Theta$  in the complete process. According to (8) and (16) the corresponding increase in black hole entropy is  $\Delta S_{bh} = \Delta\mu b\hbar^{-1} \ln 2$  (angular momentum is not added to the hole; see Appendix C). But since  $b > \pi\hbar/2T$  we have that

$$\Delta S_{bh} > (\frac{1}{2}\pi \ln 2)\Delta\mu/T. \quad (36)$$

On the other hand, the decrease in entropy of the box is clearly  $\Delta\mu/T$  (heat/temperature). Thus

$$\Delta S_c = -\Delta\mu/T. \quad (37)$$

From (36) and (37) it follows that  $\Delta(S_{bh} + S_c) > 0$  as required by the generalized second law.

#### ACKNOWLEDGMENTS

The author is grateful to Professor J. A. Wheeler for his interest in this work, his encouragement and his penetrating criticism. He also thanks Professor Karel Kuchař and Professor Brandon Carter, and Dr. Bahram Mashhoon for valuable suggestions and comments.

#### APPENDIX A

Here we shall calculate the minimum possible increase in black-hole area which must result when a spherical particle of rest mass  $\mu$  and proper radius  $b$  is captured by a Kerr black hole. We are interested in the increase in area ascribable to the particle itself, as contrasted with any

increase incidental to the process of bringing the particle to black-hole horizon. For example, there is some *circumstantial* evidence for believing that when the particle is lowered into the black hole by a string, there occurs an increase in black-hole area even as the particle is being lowered.<sup>11</sup> Furthermore, the area will experience an additional increase due to the gravitational waves radiated into the black hole by the string as it relaxes when the particle is dropped.<sup>11</sup> Similarly, if the particle falls freely to the horizon it emits gravitational waves into the hole even before it falls in; the amount of radiation may even be significant.<sup>24</sup> This radiation will also result in an increase in area. Here we shall ignore all these incidental effects and concentrate on the increase in area caused by the particle all by itself.

We assume that the particle is neutral so that it follows a geodesic of the Kerr geometry when falling freely. We shall employ Boyer-Lindquist coordinates for the charged Kerr metric<sup>25</sup>

$$ds^2 = g_{tt} dt^2 + 2g_{t\phi} dt d\phi + g_{\phi\phi} d\phi^2 + g_{rr} dr^2 + g_{\theta\theta} d\theta^2. \quad (A1)$$

For later reference we give  $g_{rr}$ :

$$g_{rr} = (r^2 + a^2 \cos^2 \theta) \Delta^{-1}, \quad (A2)$$

where

$$\Delta \equiv r^2 - 2Mr + a^2 + Q^2. \quad (A3)$$

The event horizon is located at  $r=r_+$  where  $r_+$  are defined by (7). We have

$$\Delta = (r - r_-)(r - r_+). \quad (A4)$$

First integrals for geodesic motion in the Kerr background have been given by Carter.<sup>25</sup> Christodoulou<sup>5</sup> uses the first integral

$$\begin{aligned} E^2 [r^4 + a^2(r^2 + 2Mr - Q^2)] - 2E(2Mr - Q^2)ap_\phi \\ - (r^2 - 2Mr + Q^2)p_\phi^2 - (\mu^2 r^2 + q)\Delta = (p_r \Delta)^2 \end{aligned} \quad (A5)$$

as a starting point of his analysis. In (A5)  $E = -p_r$  is the conserved energy,  $p_\phi$  is the conserved component of angular momentum in the direction of the axis of symmetry,  $q$  is Carter's fourth constant of the motion,<sup>25</sup>  $\mu$  is the rest mass of the particle, and  $p_r$  is its covariant radial momentum.

Following Christodoulou we solve (A5) for  $E$ :

$$\begin{aligned} E = & B a p_\phi + \{ [B^2 a^2 + A^{-1} (r^2 - 2Mr + Q^2)] p_\phi^2 \\ & + A^{-1} [(\mu^2 r^2 + q) \Delta + (p_r \Delta)^2] \}^{1/2}, \end{aligned} \quad (\text{A6})$$

where

$$A \equiv r^4 + a^2 (r^2 + 2Mr - Q^2), \quad (\text{A7})$$

$$B \equiv (2Mr - Q^2) A^{-1}. \quad (\text{A8})$$

At the event horizon  $\Delta = 0$  [see (A4)] so that there

$$\begin{aligned} A = & A_+ = (r_+^2 + a^2)^2, \\ B = & B_+ = (r_+^2 + a^2)^{-1}. \end{aligned} \quad (\text{A9})$$

Furthermore, at the horizon  $Ba = \Omega$  [see (9b)], and the coefficients of  $p_\phi^2$  and  $\mu^2 r^2 + q$  in (A6) vanish. However,

$$p_r \Delta = (r^2 + a^2 \cos^2 \theta) p_r$$

does not vanish at the horizon in general. If the particle's orbit intersects the horizon, then we have from (A6) that

$$E = \Omega p_\phi + A_+^{-1/2} |p_r \Delta|_+.$$

As a result of the capture, the black hole's mass increases by  $E$  and its component of angular momentum in the direction of the symmetry axis increases by  $p_\phi$ . Therefore, according to (8) the black hole's rationalized area will increase by  $\Theta^{-1} A_+^{-1/2} |p_r \Delta|_+$ . As pointed out by Christodoulou this increase vanishes only if the particle is captured from a turning point in its orbit in which case  $|p_r \Delta|_+ = 0$ . In this case we have

$$E = \Omega p_\phi. \quad (\text{A10})$$

The above analysis shows that it is possible for a black hole to capture a *point* particle without increasing its area. How is this conclusion changed if the particle has a nonzero proper radius  $b$ ? First we note that regardless of the manner in which the particle arrives at the horizon (being lowered by a string, splitting off from a second particle which then escapes, etc.), it must clearly acquire its parameters  $E$ ,  $p_\phi$ , and  $q$  while every part of it is still outside the horizon, i.e., while it is not yet part of the black hole. Moreover, as the particle is captured, it must already be detached from whatever system brought it to the horizon, so that it may be regarded as falling freely. Therefore, Eq. (A6) should always de-

scribe the motion of the particle's center of mass at the moment of capture.

It should be clear that to generalize Christodoulou's result to the present case one should evaluate (A6) not at  $r = r_+$ , but at  $r = r_+ + \delta$ , where  $\delta$  is determined by

$$\int_{r_+}^{r_+ + \delta} (g_{rr})^{1/2} dr = b$$

( $r = r_+ + \delta$  is a point a proper distance  $b$  outside the horizon). Using (A2) we find

$$b = 2\delta^{1/2} (r_+^2 + a^2 \cos^2 \theta)^{1/2} (r_+ - r_-)^{-1/2}. \quad (\text{A11})$$

To obtain this we have assumed that  $r_+ - r_- \gg \delta$  (black hole not nearly extreme). Expanding the argument of the square root in (A6) in powers of  $\delta$ , replacing  $\delta$  by its value given by (A11), and keeping only terms to  $O(b)$  we get

$$\begin{aligned} E = & \Omega p_\phi + [(r_+^2 - a^2)(r_+^2 + a^2)^{-1} p_\phi^2 + \mu^2 r_+^2 + q]^{1/2} \\ & \times \frac{1}{2} b (r_+ - r_-) (r_+^2 + a^2)^{-1} (r_+^2 + a^2 \cos^2 \theta)^{-1/2} \end{aligned} \quad (\text{A12})$$

Here we have already assumed that the particle reaches a turning point as it is captured since we know that this minimizes the increase in black-hole area. Equation (A12) is the generalization to  $O(b)$  of the Christodoulou condition (A10).

What is  $q$  in (A12)? We can obtain a lower bound for it as follows. From the requirement that the  $\theta$  momentum  $p_\theta$  be real it follows that<sup>25</sup>

$$q \geq \cos^2 \theta [a^2 (\mu^2 - E^2) + p_\phi^2 / \sin^2 \theta]; \quad (\text{A13})$$

the equality holds when  $p_\theta = 0$  at the point in question. If we replace  $E$  in (A13) by  $\Omega p_\phi$  [see (A12)] we obtain

$$q \geq \cos^2 \theta [a^2 \mu^2 + p_\phi^2 (1 / \sin^2 \theta - a^2 \Omega^2)],$$

which is correct to zeroth order in  $b$ . We know that  $1 / \sin^2 \theta \geq 1$ ; it is easily shown that  $a^2 \Omega^2 \leq \frac{1}{4}$  for a charged Kerr black hole. Therefore  $q \geq a^2 \mu^2 \cos^2 \theta$ . Substituting this into (A12) we find

$$E \geq \Omega p_\phi + \frac{1}{2} \mu b (r_+ - r_-) (r_+^2 + a^2)^{-1} \quad (\text{A14})$$

which is correct to  $O(b)$ . By retracing our steps we see that the equality sign in (A14) corresponds to the case  $p_\phi = p_\theta = p_r = 0$  at the point of capture. The increase in black-hole area, computed by means of (8), (9a), and (A14), is

$$\Delta\alpha \geq 2\mu b. \quad (\text{A15})$$

This gives the fundamental lower bound on the increase in black-hole area. We note that it is independent of  $M$ ,  $Q$ , and  $L$ .

## APPENDIX B

Here we shall calculate the minimum possible increase in black-hole area which must result when a light beam of energy  $E > 0$  coming from infinity is captured by a Kerr black hole. If the black hole is nonrotating the increase is simply obtained by setting  $dM=E$  in (8):

$$\Delta\alpha = 8ME \text{ for } a=0. \quad (\text{B1})$$

If the black hole is rotating,  $\Delta\alpha$  can be minimized by maximizing the angular momentum  $p_\phi$  which is brought in by the beam [see (8)]. To accomplish this we consider the effective potential  $V$  for the motion of a massless particle in a Kerr background.

This  $V$  is just the value of  $E$  given by (A6) regarded as a function of  $r$  for  $\mu=0$  and  $p_r=0$  ( $E$  equals  $V$  at a turning point). This potential starts off at a value  $\Omega p_\phi$  at  $r=r_+$  (see Appendix A), increases with  $r$ , reaches a maximum, and then falls off to zero as  $r\rightarrow\infty$ . For the beam to be captured by the hole it is necessary that  $p_\phi$  be small enough for the peak of the potential barrier to be smaller than  $E$  of the beam. The optimum case we seek corresponds to the peak being just equal to  $E$  so that  $p_\phi$  has its largest possible value.

It is clear that we must take  $q$  in (A6) as small as possible in order to have the lowest possible potential peak for given  $p_\phi$ . Let us first take  $q < 0$ . Then according to Carter<sup>25</sup> there are solutions to the geodesic equation only if  $|p_\phi| < aE$ . From (8) it follows that

$$\begin{aligned} \Delta\alpha &= \Theta^{-1}(E - \Omega p_\phi) \\ &> \Theta^{-1}E(1 - \Omega a). \end{aligned}$$

But since  $\Omega a \leq \frac{1}{2}$  and  $\Theta^{-1} \geq 8M$  it follows that

$$\Delta\alpha > 4ME \text{ for } q < 0. \quad (\text{B2})$$

Next we take  $q=0$ . Two cases are possible<sup>25</sup>: Either  $|p_\phi| < aE$  as above so that (B2) is again applicable, or else the orbit is purely equatorial. In the second case one may calculate the peak of the barrier numerically and then find the optimum increase in  $\alpha$ . It turns out that  $(\Delta\alpha)_{\min}$  decreases

monotonically with  $a$  for fixed  $M$ . The limit of  $(\Delta\alpha)_{\min}$  as  $a \rightarrow M$  may be computed analytically because in this limit one can find the height of the potential analytically with sufficient accuracy. One finds

$$(\Delta\alpha)_{\min} \rightarrow 4(1 - \frac{1}{2}\sqrt{3})ME \text{ as } a \rightarrow M. \quad (\text{B3})$$

It is clear that for  $q > 0$  the potential peak will be higher and the increase in area will be larger than the one given by (B3). Thus we find that the minimum increase in area results when the beam is captured by an extreme Kerr black hole from a purely equatorial orbit.

## APPENDIX C

Here we compute the value of the energy (as measured from infinity) of a particle of rest mass  $\mu$  and proper radius  $b$  which is hanging from a string just outside the horizon of a Kerr black hole. It is clear that the particle will not be moving in the  $r$  or  $\theta$  directions; hence  $p' = p_\theta = 0$  for it. We cannot claim that the particle does not move in the  $\phi$  direction. In fact, since it will be within the ergosphere in general, it cannot avoid moving in the  $\phi$  direction.<sup>1</sup> Our intuitive notion that the particle is "not moving" must be applied only in a locally nonrotating (Bardeen) frame.<sup>26</sup> The particle is at rest in such a frame if for it

$$\frac{d\phi}{dt} = -\frac{g_{t\phi}}{g_{\phi\phi}}.$$

It follows that

$$p_\phi = \mu \left( g_{t\phi} \frac{dt}{d\tau} + g_{\phi\phi} \frac{d\phi}{d\tau} \right) = 0.$$

If the particle were to be dropped, it would clearly keep its energy  $E$  and it would still have  $p_\phi = p' = p_\theta = 0$ , at least momentarily. We may thus compute  $E$  for the particle hanging in the string at a proper distance  $b$  from the horizon by setting  $p_\phi = 0$  in (A12). For  $q$  we take the value given by (A13) with the equality sign ( $p_\theta = 0$ )  $p_\phi = 0$ , and  $E = 0$  [since for  $p_\phi = 0$ ,  $E$  is of  $O(b)$ ]. Thus

$$q = \mu^2 a^2 \cos^2 \theta$$

and

$$\begin{aligned} E &= \frac{1}{2} \mu b (r_+ - r_-) (r_+^2 + a^2)^{-1} \\ &= 2\mu b \Theta. \end{aligned} \quad (\text{C1})$$

\*Based in part on the author's Ph.D. thesis, Princeton University, 1972; work supported in part by the National Science Foundation Grants No. GP 30799X to Princeton University and No. GP 32039 to the University of Texas at Austin.

†National Science Foundation Predoctoral Fellow when this work was initiated.

‡Present address.

<sup>1</sup>C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).

<sup>2</sup>R. Penrose and R. M. Floyd, *Nature* 229, 177 (1971).

<sup>3</sup>R. Penrose, *Riv. Nuovo Cimento* 1, 252 (1969).

<sup>4</sup>D. Christodoulou, *Phys. Rev. Letters* 25, 1596 (1970).

<sup>5</sup>D. Christodoulou, Ph.D. thesis, Princeton University, 1971 (unpublished).

<sup>6</sup>D. Christodoulou and R. Ruffini, *Phys. Rev. D* 4, 3552 (1971).

<sup>7</sup>S. W. Hawking, *Phys. Rev. Letters* 26, 1344 (1971); contribution to *Black Holes*, edited by B. DeWitt and C. DeWitt (Gordon and Breach, New York, 1973).

<sup>8</sup>J. M. Greif, Junior thesis, Princeton University, 1969 (unpublished).

<sup>9</sup>B. Carter, *Nature* 238, 71 (1972).

<sup>10</sup>Ya. B. Zel'dovich and I. D. Novikov, *Stars and Relativity* (University of Chicago Press, Chicago, 1971), p. 268.

<sup>11</sup>J. D. Bekenstein, Ph.D. thesis, Princeton University, 1972 (unpublished).

<sup>12</sup>J. D. Bekenstein, *Lett. Nuovo Cimento* 4, 737 (1972).

<sup>13</sup>R. Geroch, Colloquium at Princeton University, December 1971.

<sup>14</sup>We use units with  $G = c = 1$  unless otherwise specified.

<sup>15</sup>L. D. Landau and E. M. Lifshitz, *Statistical Physics* (Addison-Wesley, Reading, Mass., 1969).

<sup>16</sup>L. D. Landau and E. M. Lifshitz, *Electrodynamics of Continuous Media* (Addison-Wesley, Reading, Mass., 1960), p. 5.

<sup>17</sup>The mathematical definition of information was first given by C. E. Shannon; the relevant papers are re-

printed in C. E. Shannon and W. Weaver, *The Mathematical Theory of Communications* (University of Illinois Press, Urbana, 1949). An elementary introduction to information theory is given by M. Tribus and E. C. McIrvine, *Sci. Amer.* 225 (No. 3), 179 (1971). The derivation of statistical mechanics from information theory was first carried out by E. T. Jaynes, *Phys. Rev.* 106, 620 (1957); 108, 171 (1957).

<sup>18</sup>The relation between information theory and thermodynamics is discussed in detail by L. Brillouin, *Science and Information Theory* (Academic, New York, 1956), especially Chaps. 1, 9, 12–14.

<sup>19</sup>This was first pointed out to the author by J. A. Wheeler (private communication).

<sup>20</sup>J. D. Bekenstein, *Phys. Rev. D* 7, 949 (1973).

<sup>21</sup>See for example, A. M. Volkov, A. A. Izmest'ev, and G. V. Skrotskii, *Zh. Eksp. Teor. Fiz.* 59, 1254 (1970) [Sov. Phys. JETP 32, 686 (1971)].

<sup>22</sup>See E. Hisdal, *Phys. Norv.* 5, 1 (1971), and references cited therein.

<sup>23</sup>In Ref. 12 we gave an alternate resolution of the paradox posed by Geroch based on the apparent tendency of the black-hole area to increase as the box is being lowered.<sup>11</sup> The present approach is preferable in that it is independent of the validity of the interpretation given in Ref. 11, and in that it fits very well into the thermodynamic approach of this paper as will be evident presently.

<sup>24</sup>See M. Davis, R. Ruffini, and J. Tiomno, *Phys. Rev. D* 5, 2932 (1972) for the radiation of a particle falling radially into a Schwarzschild black hole. It is not clear whether the large amount of radiation found to go down the black hole is a device of the linearized approximation, or whether the effect will persist for other types of orbits or for Kerr black holes. Therefore we do not base any of our arguments here on this effect as we did in Refs. 11 and 12.

<sup>25</sup>B. Carter, *Phys. Rev.* 174, 1559 (1968).

<sup>26</sup>J. M. Bardeen, *Astrophys. J.* 161, 103 (1970).

## Ultraviolet Behavior of Non-Abelian Gauge Theories\*

David J. Gross † and Frank Wilczek

Joseph Henry Laboratories, Princeton University, Princeton, New Jersey 08540

(Received 27 April 1973)

It is shown that a wide class of non-Abelian gauge theories have, up to calculable logarithmic corrections, free-field-theory asymptotic behavior. It is suggested that Bjorken scaling may be obtained from strong-interaction dynamics based on non-Abelian gauge symmetry.

Non-Abelian gauge theories have received much attention recently as a means of constructing unified and renormalizable theories of the weak and electromagnetic interactions.<sup>1</sup> In this note we report on an investigation of the ultraviolet (UV) asymptotic behavior of such theories. We have found that they possess the remarkable feature, perhaps unique among renormalizable theories, of asymptotically approaching free-field theory. Such asymptotically free theories will exhibit, for matrix elements of currents between on-mass-shell states, Bjorken scaling. We therefore suggest that one should look to a non-Abelian gauge theory of the strong interactions to provide the explanation for Bjorken scaling, which has so far eluded field-theoretic understanding.

The UV behavior of renormalizable field theories can be discussed using the renormalization-group equations,<sup>2,3</sup> which for a theory involving one field (say  $g\varphi^4$ ) are

$$[m^2/\partial m + \beta(g)\partial/\partial g - n\gamma(g)]\Gamma_{\text{asy}}^{(n)}(g; P_1, \dots, P_n) = 0. \quad (1)$$

$\Gamma_{\text{asy}}^{(n)}$  is the asymptotic part of the one-particle-irreducible renormalized  $n$ -particle Green's function,  $\beta(g)$  and  $\gamma(g)$  are finite functions of the renormalized coupling constant  $g$ , and  $m$  is either the renormalized mass or, in the case of massless particles, the Euclidean momentum at which the theory is renormalized.<sup>4</sup> If we set  $P_i = \lambda q_i^0$ , where  $q_i^0$  are (nonexceptional) Euclidean momenta, then (1) determines the  $\lambda$  dependence of  $\Gamma^{(n)}$ :

$$\Gamma^{(n)}(g; P_i) = \lambda^D \Gamma^{(n)}(\bar{g}(g, t); q_i) \exp[-n \int_0^t \gamma(\bar{g}(g, t')) dt'], \quad (2)$$

where  $t = \ln \lambda$ ,  $D$  is the dimension (in mass units) of  $\Gamma^{(n)}$ , and  $\bar{g}$ , the invariant coupling constant, is the solution of

$$d\bar{g}/dt = \beta(\bar{g}), \quad \bar{g}(g, 0) = g. \quad (3)$$

The UV behavior of  $\Gamma^{(n)}$  ( $\lambda \rightarrow +\infty$ ) is determined by the large- $t$  behavior of  $\bar{g}$  which in turn is controlled by the zeros of  $\beta$ :  $\beta(g_f) = 0$ . These fixed points of the renormalization-group equations are said to be UV stable [infrared (IR) stable] if  $\bar{g} \rightarrow g_f$  as  $t \rightarrow +\infty$  ( $-\infty$ ) for  $\bar{g}(0)$  near  $g_f$ . If the physical coupling constant is in the domain of attraction of a UV-stable fixed point, then

$$\Gamma^{(n)}(g; P_i) \underset{\lambda \rightarrow \infty}{\approx} \lambda^{D-n\gamma(g_f)} \Gamma^{(n)}(g_f; q_i) \exp\{-n \int_0^\infty [\gamma(\bar{g}(g, t)) - \gamma(g_f)] dt\}, \quad (4)$$

so that  $\gamma(g_f)$  is the anomalous dimension of the field. As Wilson has stressed, the UV behavior is determined by the theory at the fixed point ( $g = g_f$ ).<sup>5</sup>

In general, the dimensions of operators at a fixed point are not canonical, i.e.,  $\gamma(g_f) \neq 0$ . If we wish to explain Bjorken scaling, we must assume the existence of a tower of operators with canonical dimensions. Recently, it has been argued for all but gauge theories, that this can only occur if the fixed point is at the origin,  $g_f = 0$ , so that the theory is asymptotically free.<sup>6,7</sup> In that case the anomalous dimensions of all operators

vanish, one obtains naive scaling up to finite and calculable powers of  $\ln \lambda$ , and the structure of operator products at short distances is that of free-field theory.<sup>7</sup> Therefore, the existence of such a fixed point, for a theory of the strong interactions, might explain Bjorken scaling and the success of naive light-cone or parton-model relations. Unfortunately, it appears that the fixed point at the origin, which is common to all theories, is not UV stable.<sup>8,9</sup> The only exception would seem to be non-Abelian gauge theories, which hitherto have not been explored in this re-

gard.

Let us consider a Yang-Mills theory given by the Lagrangian

$$\begin{aligned} \mathcal{L} &= -\frac{1}{4}F_{\mu\nu}^a F_a^{\mu\nu}, \\ F_{\mu\nu}^a &= \partial_\mu A_\nu^a - \partial_\nu A_\mu^a - g C_{abc} A_\mu^b A_\nu^c, \end{aligned} \quad (5)$$

where the  $C_{abc}$  are the structure constants of some (semisimple) Lie group  $G$ . Since the theory is massless, the renormalization is performed at an (arbitrary) Euclidean point. For example, the wave-function renormalization constant  $Z_3(g, \Lambda/m)$  will be defined in terms of the unrenormalized vector-meson propagator  $D_{\mu\nu}^{ab}$  (in the Landau gauge),

$$D_{\mu\nu}^{ab}(P)|_{P^2=-m^2} = \left( g_{\mu\nu} + \frac{P_\mu P_\nu}{m^2} \right) \frac{iZ_3}{m^2} \delta_{ab}. \quad (6)$$

(For a thorough discussion of the renormalization see the work of Lee and Zinn-Justin.<sup>10</sup>) The renormalization-group equations for this theory are easily derived.<sup>11</sup> In the Landau gauge they are identical with (1). In order to investigate the stability of the origin, it is sufficient to calculate  $\beta$  to lowest order in perturbation theory. To this order we have

$$\beta(g) = \frac{\partial g}{\partial \ln m} \Big|_{\Lambda_{1g0}} = -g \frac{\partial}{\partial \ln \Lambda} \left( \frac{Z_3^{3/2}}{Z_1} \right), \quad (7)$$

where  $\Lambda$  is a UV cutoff, and  $Z_1$  the charge-renormalization constant. In Abelian gauge theories  $Z_3 = Z_1 = 1 - g^2 C \ln \Lambda$  ( $C > 0$ ), as a consequence of gauge invariance and the Källén-Lehman representation, and thus  $\beta(g) \cong g^3$  which leads to IR stability at  $g = 0$ . Non-Abelian theories have no such requirement;  $Z_3$  and  $Z_1$  are gauge dependent and can be greater than 1. Thus  $\beta(g)$ , which must be gauge independent in lowest order, could have any sign at  $g = 0$ . We have calculated  $Z_1$  and  $Z_3$  for the above Lagrangian, and we find that<sup>12</sup>

$$\beta_V = -(g^3/16\pi^2)^{11/3} C_2(G) + O(g^5), \quad (8)$$

where  $C_2(G)$  is the quadratic Casimir operator of the adjoint representation of the group  $G$ :  $\sum_{b,c} c_{abc} \times c_{abc} = C_2(G) \delta_{aa}$  [e.g.,  $C_2(\text{SU}(N)) = N$ ]. The solution of (3) is then  $\bar{g}^2(t) = g^2/(1 - 2\beta_V g^{-1}t)$ , and  $\bar{g} \rightarrow 0$  as  $t \rightarrow \infty$  as long as the physical coupling constant  $g$  is in the domain of attraction of the origin.<sup>13</sup>

We have thus established that for all non-Abelian gauge theories based on semisimple Lie groups the origin is UV stable. It is easy to incorporate fermions into such a theory without destroying the UV stability. The fermion interac-

tion is given by  $L_F = \bar{\psi}(i\gamma^\cdot \delta - g\gamma^\cdot B^a M^a)\psi + \text{mass terms}$ , where  $M^a$  are the matrices of some representation  $R$  of the gauge group  $G$ . The only effect of the fermions is to change the value of  $\beta(g)$  by the amount<sup>11</sup>

$$\beta_F(R) = (g^3/16\pi^2)^{4/3} T(R), \quad (9)$$

where  $\text{Tr}(M^a M^b) = T(R) \delta_{ab}$ ,  $T(R) = C_2(R) d(R)/r$ ,  $d(R)$  is the dimension of the representation  $R$ , and  $r$  is the order of the group, i.e., the number of generators, and  $C_2(R)$  is the quadratic Casimir operator of the representation. Although the fermions tend to destabilize the origin, there is room to spare. For example, in the case of  $\text{SU}(3)$ :  $\beta_V = -11$ , whereas  $\beta_F(3) = \frac{2}{3}$ ,  $\beta_F(8) = 4$ , etc., so that one could accommodate as many as sixteen triplets. One can therefore construct many asymptotically free theories with fermions. The vector mesons, however, will remain massless until the gauge symmetry is spontaneously broken. One might hope that this would be a consequence of the dynamics,<sup>14</sup> but at the present the only known way of achieving this is to introduce scalar Higgs mesons, whose nonvanishing vacuum expectation values break the symmetry.

The introduction of scalar mesons has a very destabilizing effect on the UV stability of the origin. Their contribution to  $\beta(g)$  is small; a scalar meson transforming under a complex (real) representation  $R$  of the gauge group adds to  $\beta$  a term equal to  $\frac{1}{4} (\frac{1}{8})$  of Eq. (9). The problem with scalar mesons is that they necessarily have their own quartic couplings, and one must deal with a new coupling constant. Consider the Lagrangian for the coupling of scalars belonging to a representation  $R$  of  $G$ :

$$\mathcal{L} = \frac{1}{2} [(\partial_\mu - igB_\mu^a M^a) \vec{\varphi}]^2 - \lambda(\vec{\varphi} \cdot \vec{\varphi})^2 + V(\vec{\varphi}), \quad (10)$$

where  $V(\varphi)$  contains cubic, quadratic, and linear terms in  $\varphi$  (which have no effect on the UV behavior of the theory) plus, perhaps, additional quartic terms invariant under  $G$ . The renormalization-group equations have an additional term,  $\beta_\lambda(g, \lambda) \partial/\partial \lambda$ , and one must investigate the UV stability of the origin ( $g = \lambda = 0$ ) with respect to both  $g$  and  $\lambda$  [if there are other quartic invariants in  $V(\varphi)$  there will be additional coupling constants to consider]. The structure of the renormalization-group equation for  $g$  is unchanged to lowest order, whereas for the coupling constant  $I \equiv \lambda/g^2$  we have<sup>11</sup>

$$d\bar{\Gamma}(\Gamma, t, g^2)/dt = \bar{g}^2 [A\bar{\Gamma}^2 + B\bar{\Gamma} + C] \quad (11)$$

(where we have neglected terms of order  $g^4$ ,  $g^4\Gamma$ ,

$g^4\Gamma^2$ , and  $g^4\Gamma^3$ ). In the absence of vector mesons ( $g=0$ ) this equation is UV unstable at  $\lambda=0$ , since  $A$  is strictly positive and  $\lambda$  must be positive.<sup>15</sup> The vector mesons contribute to  $B$  and  $C$  and tend to stabilize the origin. If the right-hand side of (11) has positive zeros ( $C>0$ ,  $B<0$ , and  $B^2 > 4AC$ ), then for  $\Gamma$  less than the larger zero of (11) we will have that  $\lambda \rightarrow +0$  as  $t \rightarrow \infty$ . We have investigated the structure of these equations for a large class of gauge theories and representations of the scalar mesons. We have found many examples of theories which contain scalar mesons and are UV stable.<sup>11</sup> These include (a)  $SU(N)$  if the scalar mesons belong to the adjoint representation for  $N \geq 6$ ; (b)  $SU(N) \otimes SU(N)$  if the scalars belong to the  $(N, \bar{N})$  representation for  $N \geq 5$ ; (c)  $SU(N)$  with the scalars transforming as a symmetric tensor for  $N \geq 9$ ; and many others. In all of these models it is necessary for the theory to contain a large number of fermions in order to make  $\beta_g$  small; otherwise  $\bar{g}$  approaches zero too rapidly for the vector mesons to stabilize the scalar couplings.

Unfortunately, in none of these models can the gauge symmetry be totally broken by the Higgs mechanism. The requirement that the interactions of the scalar mesons be renormalizable so severely constrains the form of Lagrangian that the ground state invariably is invariant under some non-Abelian subgroup of the gauge group. If one tries to overcome this by larger representations for the scalar mesons, UV instability inevitably occurs.

It thus appears to be very difficult to retain UV stability and break the gauge symmetry by explicitly introducing Higgs mesons. Since the Higgs mesons are so restrictive, we would prefer to believe that spontaneous symmetry breaking would arise dynamically.<sup>14</sup> This is suggested by the IR instability of the theories, which assures us that perturbation theory is not trustworthy with respect to the stability of the symmetric theory nor to its particle content.

With this hope in mind one can construct many interesting models of the strong interactions. One particularly appealing model is based on three triplets<sup>16</sup> of fermions, with Gell-Mann's  $SU(3) \otimes SU(3)$  as a global symmetry and an  $SU(3)$  "color" gauge group to provide the strong interactions. That is, the generators of the strong-interaction gauge group commute with ordinary  $SU(3) \otimes SU(3)$  currents and mix quarks with the same isospin and hypercharge but different "color." In such a model the vector mesons are

neutral, and the structure of the operator product expansion of electromagnetic or weak currents is (assuming that the strong coupling constant is in the domain of attraction of the origin!) essentially that of the free quark model (up to calculable logarithmic corrections).<sup>11</sup>

Finally, we note that theories of the weak and electromagnetic interactions, built on semisimple Lie groups,<sup>17</sup> will be asymptotically free if we again ignore the complications due to the Higgs particles. This suggests that the program of Baker, Johnson, Willey, and Adler<sup>18</sup> to calculate the fine-structure constant as the value of the UV-stable fixed point in quantum electrodynamics might fail for such theories.

\*Research supported by the U.S. Air Force Office of Scientific Research under Contract No. F-44620-71-C-0180.

†Alfred P. Sloan Foundation Research Fellow.

<sup>1</sup>S. Weinberg, Phys. Rev. Lett. 19, 1264 (1967). For an extensive review as well as a list of references, see B. W. Lee, in Proceedings of the Sixteenth International Conference on High Energy Physics, National Accelerator Laboratory, Batavia, Illinois, 1972 (to be published).

<sup>2</sup>M. Gell-Mann and F. E. Low, Phys. Rev. 95, 1300 (1954).

<sup>3</sup>C. G. Callan, Phys. Rev. D 2, 1541 (1970); K. Symanzik, Commun. Math. Phys. 18, 227 (1970).

<sup>4</sup>The basic assumption underlying the derivation and utilization of the renormalization group equations is that the large Euclidean momentum behavior of the theory is the same as the sum, to all orders, of the leading powers in perturbation theory.

<sup>5</sup>K. Wilson, Phys. Rev. D 3, 1818 (1971).

<sup>6</sup>G. Parisi, to be published.

<sup>7</sup>C. G. Callan and D. J. Gross, to be published.

<sup>8</sup>A. Zee, to be published.

<sup>9</sup>S. Coleman and D. J. Gross, to be published.

<sup>10</sup>B. W. Lee and J. Zinn-Justin, Phys. Rev. D 5, 3121 (1972).

<sup>11</sup>Full details will be given in a forthcoming publication: D. J. Gross and F. Wilczek, to be published.

<sup>12</sup>After completion of this calculation we were informed of an independent calculation of  $\beta$  for gauge theories coupled to fermions by H. D. Politzer [private communication, and following Letter, Phys. Rev. Lett. 30, 1346 (1973)].

<sup>13</sup>K. Wilson has suggested that the coupling constants of the strong interactions are determined to be IR-stable fixed points. For nongauge theories the IR stability of the origin in four-dimensional field theories implies that theories so constructed are trivial, at least in a domain about the origin. Our results suggest that non-Abelian gauge theories might possess IR-stable fixed points at nonvanishing values of the coupling constants.

<sup>14</sup>Y. Nambu and G. Jona-Lasinio, Phys. Rev. 122, 345 (1961); S. Coleman and E. Weinberg, Phys. Rev. D 7, 1888 (1973).

<sup>15</sup>K. Symanzik (to be published) has recently suggested that one consider a  $\lambda\phi^4$  theory with a negative  $\lambda$  to achieve UV stability at  $\lambda=0$ . However, one can show, using the renormalization-group equations, that in such theory the ground-state energy is unbounded from below (S. Coleman, private communication).

<sup>16</sup>W. A. Bardeen, H. Fritzsch, and M. Gell-Mann, CERN Report No. CERN-TH-1538, 1972 (to be published).

<sup>17</sup>H. Georgi and S. L. Glashow, Phys. Rev. Lett. 28, 1494 (1972); S. Weinberg, Phys. Rev. D 5, 1962 (1972).

<sup>18</sup>For a review of this program, see S. L. Adler, in Proceedings of the Sixteenth International Conference on High Energy Physics, National Accelerator Laboratory, Batavia, Illinois, 1972 (to be published).

## Unity of All Elementary-Particle Forces

Howard Georgi\* and S. L. Glashow

*Lyman Laboratory of Physics, Harvard University, Cambridge, Massachusetts 02138*

(Received 10 January 1974)

Strong, electromagnetic, and weak forces are conjectured to arise from a single fundamental interaction based on the gauge group  $SU(5)$ .

We present a series of hypotheses and speculations leading inescapably to the conclusion that  $SU(5)$  is the gauge group of the world—that all elementary particle forces (strong, weak, and electromagnetic) are different manifestations of the same fundamental interaction involving a single coupling strength, the fine-structure constant. Our hypotheses may be wrong and our speculations idle, but the uniqueness and simplicity of our scheme are reasons enough that it be taken seriously.

Our starting point is the assumption that *weak and electromagnetic forces are mediated by the vector bosons of a gauge-invariant theory with spontaneous symmetry breaking*. A model describing the interactions of leptons using the gauge group  $SU(2) \otimes U(1)$  was first proposed by Glashow, and was improved by Weinberg and Salam who incorporated spontaneous symmetry breaking.<sup>1</sup> This scheme can also describe hadrons, and is just one example of an infinite class of models compatible with observed weak-interaction phenomenology. If we assume that *there are as few fermion fields as possible* and, in particular, that there are no unobserved leptons, the Weinberg model becomes unique up to extensions of the gauge group: The observed leptons may be described by six left-handed Weyl fields ( $e_L^-$ ,  $\mu_L^-$ ,  $\nu_L$ ,  $\nu_L'$ ,  $e_L^+$ ,  $\mu_L^+$ ) and their charge conjugates. If the gauge couplings do not mix leptons with quarks, these six fields must transform as a representation of the gauge group: one of the 23 subgroups of  $U(6)$  containing an  $SU(2) \otimes U(1)$  subgroup in which the leptons behave as they do in the Weinberg model.

To include hadrons in the theory, we must use the Glashow-Iliopoulos-Maiani (GIM) mechanism and introduce a fourth quark  $p'$  carrying charm.<sup>2</sup> Still, decisions must be made: Should the quarks have fractional or integer charges? Should there be one quartet of quarks or several? Bouchiat, Iliopoulos, and Meyer suggested what seems the most attractive alternative: *three quartets of fractionally charged quarks*.<sup>3</sup> This combination

of the GIM mechanism with the notion of colored quarks<sup>4</sup> keeps the successes of the quark model and gives an important bonus: Lepton and hadron anomalies cancel so that the theory of weak and electromagnetic interactions is renormalizable.<sup>5</sup>

The next step is to include strong interactions. We assume that *strong interactions are mediated by an octet of neutral vector gauge gluons* associated with local color  $SU(3)$  symmetry, and that there are no fundamental strongly interacting scalar-meson fields.<sup>6</sup> This insures that parity and hypercharge are conserved to order  $\alpha$ ,<sup>7</sup> and does not lead to any new anomalies, so that the theory remains renormalizable. The strongest binding forces are in color singlet states which may explain why observed hadrons lie in  $qqq$  and  $q\bar{q}$  configurations.<sup>8</sup> And, it gives another important bonus: Since the strong interactions are associated with a non-Abelian theory, they may be asymptotically free.<sup>9</sup>

Thus, we see how attractive it is for strong, weak, and electromagnetic interactions to spring from a gauge theory based on the group  $\mathcal{F} = SU(3) \otimes SU(2) \otimes U(1)$ . Alas, this theory is defective in one important respect: It does not truly unify weak and electromagnetic interactions. The  $SU(2) \otimes U(1)$  gauge couplings describe two interactions with two independent coupling constants; a true unification would involve only one.

Electric charge is observed to be quantized. This has no natural explanation in the framework of conventional quantum electrodynamics, but it is necessarily true in any unified theory<sup>10</sup>—yet another reason to search for a true unification.

We must assume that the gauge group is larger than  $\mathcal{F}$ . Suppose it is of the form  $SU(3) \otimes \mathcal{W}$  where  $\mathcal{W}$  contains  $SU(2) \otimes U(1)$  but has a unique gauge coupling constant.  $\mathcal{W}$  must be simple, or the direct product of isomorphic simple factors with discrete symmetries which interchange them. This embedding of the Weinberg model implies a relationship between the coupling constants of the  $SU(2)$  and  $U(1)$  subgroups. Because leptons are singlets under color  $SU(3)$ , leptons and quarks

must lie in separate representations of  $\mathfrak{W}$ . If only the six observed lepton states are involved,  $\mathfrak{W}$  must be one of the 23 relevant subgroups of  $U(6)$ . The only candidates involving a single gauge coupling constant are  $SU(3)$ ,  $SU(3) \otimes SU(3)$ , and  $SU(6)$ .<sup>11</sup> For each of these cases, the mixing angle is fixed so that  $\sin^2\theta_w = \frac{1}{4}$ . None of these schemes can describe hadrons: The generator corresponding to electric charge does not admit fractional charges, nor, being traceless, can it explain why the sum of the quark charges is not zero. No gauge group of the form  $SU(3) \otimes \mathfrak{W}$  works.

We see that we cannot unify weak and electromagnetic interactions independently of strong interactions. The remaining possibility is that the gauge group  $\mathfrak{G}$  contains  $\mathfrak{F}$  as a subgroup but is itself simple or the direct product of isomorphic simple factors. Leptons and quarks must lie together in the same irreducible representations of such a group: Some gauge fields carry lepton number and quark number. The same coupling strength—the fine-structure constant—characterizes all three kinds of interaction. This outrageous possibility may seem palatable after the following discussion about asymptotic freedom and its complement, infrared slavery.

Asymptotic freedom is a property of non-Abelian Yang-Mills field theories which promises to explain the pointlike structure of hadrons at high energy.<sup>9</sup> Unfortunately, these theories do not appear to describe strong interactions correctly since they involve massless strongly interacting vector bosons. The obvious solution is to introduce strongly interacting scalar mesons which develop vacuum expectation values, spontaneously break the gauge symmetry, and generate vector meson masses by the Higgs mechanism. But the scalar-meson Lagrangian involves additional renormalizable couplings which may spoil asymptotic freedom. Sadly, no one has found an asymptotically free model in which the gauge symmetry is completely broken and all the vector mesons develop mass.<sup>12</sup>

Weinberg, and Gross and Wilczek,<sup>13</sup> propose an astonishingly radical solution: to leave the gauge symmetry unbroken. While the Yang-Mills Lagrangian appears to describe massless vector bosons, the hideous infrared divergences of the theory conspire to prevent their appearance in physical states. This could explain the absence of physical states that are not color singlets and answer the old saw: Why don't the quarks get out? We have nothing to say about the merits of

this picture; we assume that it works and use it.

The essential thing about a theory of strong interactions based on an unbroken non-Abelian gauge symmetry is that the strength of strong interactions no longer depends on the existence of a large coupling constant. Even if the gauge coupling constant is small, say of order  $e$ , the infrared divergences of the theory can lead to phenomenological interactions strong enough to keep the quarks bound.<sup>14</sup> *What we want is not asymptotic freedom but infrared slavery.*

The theory we have in mind involves a unifying gauge group  $\mathfrak{G}$  whose only coupling constant is the unit of electric charge and which contains—in an appropriate way—the subgroup  $\mathfrak{F}$ . The symmetry is spontaneously broken leaving only the direct product of color  $SU(3)$  and electromagnetic gauge invariance as exact local symmetries. Color  $SU(3)$  is an unbroken non-Abelian gauge symmetry causing infrared slavery and leading to strong interactions. Electromagnetic gauge invariance is Abelian and commutes with color  $SU(3)$ . Since the photon has no direct couplings to the gauge fields of color  $SU(3)$ , electromagnetism is free of insolvable infrared-divergence problems, and photons may be freely emitted and absorbed. All other gauge fields develop masses through the Higgs mechanism. Those associated with the subgroup  $SU(3) \otimes U(1)$ , aside from the photon, mediate ordinary weak interactions and the neutral-current effects of the Weinberg model. The rest, which are colored and massive, mediate new and presumably even weaker interactions.

Our unifying group  $\mathfrak{G}$  must be of rank at least 4. There are exactly nine rank-4 local Lie groups which can involve only one coupling strength:  $[SU(2)]^4$ ,  $[O(5)]^2$ ,  $[SU(3)]^2$ ,  $[G_2]^2$ ,  $O(8)$ ,  $O(9)$ ,  $Sp(8)$ ,  $F_4$ , and  $SU(5)$ .<sup>1</sup> The first two are unacceptable since they do not contain  $SU(3)$ . To proceed, we review the behavior of quarks and leptons under  $\mathfrak{F}$ .

We use the Weyl notation in which all fermion fields are left-handed two-component spinors. There are thirty such fields in our picture of nature: four leptons  $(\mu^-, \nu', e^-, \nu)_L$ , two antileptons  $(\mu^+, e^+)_L$ , twelve quarks  $(p_i', p_i, n_i, \lambda_i)_L$ , and twelve antiquarks  $(\bar{p}_i', \bar{p}_i, \bar{n}_i, \bar{\lambda}_i)_L$ , where the color index  $i$  assumes three values. Under the subgroup  $SU(3) \otimes SU(2)$ , the leptons are  $SU(3)$  singlets and  $SU(2)$  doublets; the antileptons are singlets under both groups; the quarks are  $SU(3)$  triplets as well as  $SU(2)$  doublets; and, finally, the anti-quarks are  $SU(3)$   $3^*$ 's but  $SU(2)$  singlets.

The  $SU(3) \otimes SU(2)$  content of the thirty fields is  $2(\underline{1}, 2) \oplus 2(\underline{1}, 1) \oplus 2(\underline{3}, 2) \oplus 4(\underline{3}^*, 1)$  in an evident notation.

This representation is complex, not equivalent to its complex conjugate. So also is the corresponding representation of  $\mathcal{F}$ . Of our nine candidates only  $[SU(3)]^2$  and  $SU(5)$  admit complex representations. We have already considered and rejected  $[SU(3)]^2$  in our discussion of the synthesis of just weak and electromagnetic interactions. We are left with  $SU(5)$ .

Under the subgroup  $SU(3) \otimes SU(2)$ , the fundamental five-dimensional representation of  $SU(5)$  transforms like  $(\underline{1}, 2) \oplus (\underline{3}, 1)$ . The complex conjugate  $\underline{5}^*$  transforms like  $(\underline{1}, 2) \oplus (\underline{3}^*, 1)$ . The irreducible ten-dimensional representation given by the antisymmetrized tensor product of two  $\underline{5}$ 's transforms like  $(\underline{1}, 1) \oplus (\underline{3}^*, 1) \oplus (\underline{3}, 2)$ . If the thirty left-handed fermions transform like two  $\underline{10}$ 's and two  $\underline{5}^*$ 's, the  $\mathcal{F}$  content is just right to describe physics. In order to display these representations, we replace the two  $\underline{5}^*$ 's of left-handed fields by two  $\underline{5}$ 's of their right-handed charge conjugates. The representations containing electrons are then a  $\underline{5}$  and a  $\underline{10}$ :

$$\begin{bmatrix} n_1 \\ n_2 \\ n_3 \\ e^+ \\ \nu \end{bmatrix}_R, \quad \frac{1}{\sqrt{2}} \begin{bmatrix} 0 & \bar{p}_3 & -\bar{p}_2 & -p_1(\theta) & -n_1 \\ -\bar{p}_3 & 0 & \bar{p}_1 & -p_2(\theta) & -n_2 \\ \bar{p}_2 & -p_1 & 0 & -p_3(\theta) & -n_3 \\ p_1(\theta) & p_2(\theta) & p_3(\theta) & 0 & -e^+ \\ n_1 & n_2 & n_3 & e^+ & 0 \end{bmatrix}_L$$

where  $p(\theta) = p \cos\theta - p' \sin\theta$ . The  $\underline{5}$  and  $\underline{10}$  containing muons are obtained from these by the replacements  $e^+ \rightarrow \mu^+$ ,  $\nu \rightarrow \nu'$ ,  $n \rightarrow \lambda$ ,  $\bar{p} \rightarrow \bar{p}'$ , and  $p(\theta) \rightarrow p'(\theta) = p' \cos\theta + p \sin\theta$ .

Having the representations before us, we answer the obvious questions: Are there anomalies? What Higgs mesons are necessary? What mixing angle is predicted? What new interactions are predicted?

While we already know that the  $\mathcal{F}$  subgroup is free of anomalies for the representation we have chosen, the full unifying group might not be. But it is! Remarkably, the  $\underline{5}^*$  and  $\underline{10}$  have equal and opposite anomalies: Our theory is entirely anomaly free. Indeed,  $SU(5)$  is the only group of any rank with a thirty-dimensional, anomaly-free representation with the correct  $\mathcal{F}$  content.

Two irreducible representations of Higgs mesons are needed. We need a multiplet with a very large vacuum expectation value to break the  $SU(5)$  symmetry down to  $\mathcal{F}$ . This is done most simply with 24 real scalar-meson fields transforming like the adjoint representation. It is the

analog to the superstrong breaking discussed by Weinberg in his treatment of  $SU(3) \otimes SU(3)$ .<sup>11</sup> All the vector bosons except the twelve associated with generators of  $\mathcal{F}$  develop superheavy masses and can hopefully be neglected. We also need Higgs mesons to give mass to the fermions and the weak-interaction intermediaries. For the most general zeroth-order mass matrix consistent with exact color  $SU(3)$  symmetry, we need five complex scalar-meson fields transforming like the fundamental representation and 45 complex scalar-meson fields transforming like the  $\underline{45}$  contained in  $\underline{5}^* \times \underline{10}$ . If only the  $\underline{5}$  is present, the  $p$  and  $p'$  masses and the Cabibbo angle are arbitrary, but the other masses satisfy the relations  $m_n = m_e$  and  $m_\lambda = m_\mu$ . Does this mean that the muon-electron mass splitting has the same origin as  $SU(3)$  breaking?

For the mixing angle, the theory predicts  $\sin^2 \theta_w = \frac{3}{8}$ .

Finally we come to a discussion of superweak interactions and  $SU(3)$ -colored superheavy vector bosons. In addition to mediating such bizarre interactions as  $K^0 \rightarrow \mu^+ e^-$ , they make the proton unstable. For instance, there is a superheavy colored vector boson which causes the virtual transitions  $p_1 + p_2 \rightarrow W \rightarrow \bar{n}_3 + e^+$ . Exchange of this vector boson contributes directly to the decay  $p \rightarrow \pi^0 + e^+$ . Since the proton is rather stable,<sup>15</sup> this vector boson must be very massive.<sup>16</sup> The Higgs mesons can also mediate proton decay, and must also be very massive.

From simple beginnings we have constructed the unique simple theory. It makes just one easily testable prediction,  $\sin^2 \theta_w = \frac{3}{8}$ . It also predicts that the proton decays—but with an unknown and adjustable rate. More theoretical work is needed to determine whether the idea of infrared slavery, necessary for our unification, actually makes sense.

\*Junior Fellow, Harvard University, Society of Fellows. Work supported in part by the U.S. Air Force Office of Scientific Research under Contract No. F44620-70-C-0030 and by the National Science Foundation under Grant No. GP-3081X.

<sup>1</sup>S. L. Glashow, Nucl. Phys. **22**, 579 (1961); S. Weinberg, Phys. Rev. Lett. **19**, 1264 (1967); A. Salam, in *Proceedings of the Eighth Nobel Symposium, on Elementary Particle Theory, Relativistic Groups, and Analyticity, Stockholm, Sweden, 1968*, edited by N. Svartholm (Almqvist and Wiksell, Stockholm, 1968).

<sup>2</sup>S. L. Glashow, J. Iliopoulos, and L. Maiani, Phys. Rev. D 2, 1285 (1970).

<sup>3</sup>C. Bouchiat, J. Iliopoulos, and Ph. Meyer, Phys. Lett. 38B, 519 (1972).

<sup>4</sup>M. Gell-Mann, Acta Phys. Austr. Suppl. IX, 733 (1972).

<sup>5</sup>The same statements could be made about a three-quartet model with integral charged quarks, in the spirit of M. Y. Han and Y. Nambu, Phys. Rev. 139, B1006 (1965). In such a scheme, electric charge does not commute with color SU(3).

<sup>6</sup>J. C. Pati and A. Salam, Phys. Rev. D 8, 1240 (1973), and Phys. Rev. Lett. 31, 661 (1973).

<sup>7</sup>S. Weinberg, Phys. Rev. D 8, 605 (1973), and Phys. Rev. Lett. 31, 494 (1973).

<sup>8</sup>O. W. Greenberg, Phys. Rev. Lett. 13, 598 (1964).

<sup>9</sup>H. D. Politzer, Phys. Rev. Lett. 30, 1346 (1973); D. J. Gross and F. Wilczek, Phys. Rev. Lett. 30, 1343 (1973); T. Appelquist and H. Georgi, Phys. Rev. D 8, 4000 (1973); A. Zee, Phys. Rev. D (to be published); H. Georgi and H. D. Politzer, Phys. Rev. D (to be published); D. J. Gross and F. Wilczek, Phys. Rev. D 8, 3633 (1973), and Phys. Rev. D (to be published).

<sup>10</sup>In a general gauge theory, the Lie algebra of the gauge group is a direct sum of a semisimple and an Abelian Lie algebra. Only if the Abelian term is absent—as in a unified theory—is the gauge group necessarily compact, and charge necessarily quantized. Assume that electric charge  $Q$  were not quantized in such a theory, i.e., that its eigenvalues were not commensurate. The topological closure of the one-param-

eter subgroup  $\{\exp(i\alpha Q)\}$  would be a compact Abelian Lie group with at least two parameters. Let its Lie algebra  $A$  be spanned by  $Q, Q_1, \dots, Q_n$ , where  $n > 0$ . Because the gauge group is compact, its Lie algebra contains  $A$  and the  $Q_i$  are associated with gauge fields other than the photon. Because  $\exp(i\alpha Q)$  is an unbroken local symmetry, the  $Q_i$  generate unbroken local symmetries and their associated gauge fields are massless. Since there is only one massless gauge field, we conclude that  $Q$  is quantized.

<sup>11</sup>S. Weinberg, Phys. Rev. D 5, 1962 (1972); H. Georgi and S. L. Glashow, Phys. Rev. D 7, 2457 (1973).

<sup>12</sup>See, for instance, T. P. Cheng, E. Eichten, and L.-F. Li, SLAC Report No. SLAC-PUB-1340, 1973 (unpublished).

<sup>13</sup>See the second paper in Ref. 7 and the sixth paper in Ref. 9.

<sup>14</sup>S. Weinberg, to be published.

<sup>15</sup>H. S. Gurn, W. R. Kropp, F. Reines, and B. Meyer, Phys. Rev. 158, 1321 (1967).

<sup>16</sup>A naive calculation indicates that the vector boson mass must be greater than  $10^{15}$  GeV  $\approx 10^{-9}$  g! Let the reader who finds this hard to swallow double the number of fermion states and put quarks and leptons in different (but equivalent) thirty-dimensional representations. He must introduce both weakly interacting quarks and strongly interacting leptons. Now quark number is conserved modulo two and the proton is stable. The deuteron decays via the exchange of four superweak vector bosons, but this is not a serious problem.

# Particle Creation by Black Holes

S. W. Hawking

Department of Applied Mathematics and Theoretical Physics, University of Cambridge,  
Cambridge, England

Received April 12, 1975

**Abstract.** In the classical theory black holes can only absorb and not emit particles. However it is shown that quantum mechanical effects cause black holes to create and emit particles as if they were hot bodies with temperature  $\frac{\hbar\kappa}{2\pi k} \approx 10^{-6} \left(\frac{M_\odot}{M}\right)^\circ\text{K}$  where  $\kappa$  is the surface gravity of the black hole. This thermal emission leads to a slow decrease in the mass of the black hole and to its eventual disappearance: any primordial black hole of mass less than about  $10^{15}$  g would have evaporated by now. Although these quantum effects violate the classical law that the area of the event horizon of a black hole cannot decrease, there remains a Generalized Second Law:  $S + \frac{1}{4}A$  never decreases where  $S$  is the entropy of matter outside black holes and  $A$  is the sum of the surface areas of the event horizons. This shows that gravitational collapse converts the baryons and leptons in the collapsing body into entropy. It is tempting to speculate that this might be the reason why the Universe contains so much entropy per baryon.

## 1.

Although there has been a lot of work in the last fifteen years (see [1, 2] for recent reviews), I think it would be fair to say that we do not yet have a fully satisfactory and consistent quantum theory of gravity. At the moment classical General Relativity still provides the most successful description of gravity. In classical General Relativity one has a classical metric which obeys the Einstein equations, the right hand side of which is supposed to be the energy momentum tensor of the classical matter fields. However, although it may be reasonable to ignore quantum gravitational effects on the grounds that these are likely to be small, we know that quantum mechanics plays a vital role in the behaviour of the matter fields. One therefore has the problem of defining a consistent scheme in which the space-time metric is treated classically but is coupled to the matter fields which are treated quantum mechanically. Presumably such a scheme would be only an approximation to a deeper theory (still to be found) in which space-time itself was quantized. However one would hope that it would be a very good approximation for most purposes except near space-time singularities.

The approximation I shall use in this paper is that the matter fields, such as scalar, electro-magnetic, or neutrino fields, obey the usual wave equations with the Minkowski metric replaced by a classical space-time metric  $g_{ab}$ . This metric satisfies the Einstein equations where the source on the right hand side is taken to be the expectation value of some suitably defined energy momentum operator for the matter fields. In this theory of quantum mechanics in curved space-time there is a problem in interpreting the field operators in terms of annihilation and creation operators. In flat space-time the standard procedure is to decompose

the field into positive and negative frequency components. For example, if  $\phi$  is a massless Hermitian scalar field obeying the equation  $\phi_{;ab}\eta^{ab}=0$  one expresses  $\phi$  as

$$\phi = \sum_i \{f_i a_i + \bar{f}_i a_i^\dagger\} \quad (1.1)$$

where the  $\{f_i\}$  are a complete orthonormal family of complex valued solutions of the wave equation  $f_{i;ab}\eta^{ab}=0$  which contain only positive frequencies with respect to the usual Minkowski time coordinate. The operators  $a_i$  and  $a_i^\dagger$  are interpreted as the annihilation and creation operators respectively for particles in the  $i$ th state. The vacuum state  $|0\rangle$  is defined to be the state from which one cannot annihilate any particles, i.e.

$$a_i|0\rangle = 0 \quad \text{for all } i.$$

In curved space-time one can also consider a Hermitian scalar field operator  $\phi$  which obeys the covariant wave equation  $\phi_{;ab}g^{ab}=0$ . However one cannot decompose into its positive and negative frequency parts as positive and negative frequencies have no invariant meaning in curved space-time. One could still require that the  $\{f_i\}$  and the  $\{\bar{f}_i\}$  together formed a complete basis for solutions of the wave equations with

$$\frac{1}{2}i \int_S (f_i \bar{f}_{j;a} - \bar{f}_j f_{i;a}) d\Sigma^a = \delta_{ij} \quad (1.2)$$

where  $S$  is a suitable surface. However condition (1.2) does not uniquely fix the subspace of the space of all solutions which is spanned by the  $\{f_i\}$  and therefore does not determine the splitting of the operator  $\phi$  into annihilation and creation parts. In a region of space-time which was flat or asymptotically flat, the appropriate criterion for choosing the  $\{f_i\}$  is that they should contain only positive frequencies with respect to the Minkowski time coordinate. However if one has a space-time which contains an initial flat region (1) followed by a region of curvature (2) then a final flat region (3), the basis  $\{f_{1i}\}$  which contains only positive frequencies on region (1) will not be the same as the basis  $\{f_{3i}\}$  which contains only positive frequencies on region (3). This means that the initial vacuum state  $|0_1\rangle$ , the state which satisfies  $a_{1i}|0_1\rangle=0$  for each initial annihilation operator  $a_{1i}$ , will not be the same as the final vacuum state  $|0_3\rangle$  i.e.  $a_{3i}|0_1\rangle \neq 0$ . One can interpret this as implying that the time dependent metric or gravitational field has caused the creation of a certain number of particles of the scalar field.

Although it is obvious what the subspace spanned by the  $\{f_i\}$  is for an asymptotically flat region, it is not uniquely defined for a general point of a curved space-time. Consider an observer with velocity vector  $v^a$  at a point  $p$ . Let  $B$  be the least upper bound  $|R_{abcd}|$  in any orthonormal tetrad whose timelike vector coincides with  $v^a$ . In a neighbourhood  $U$  of  $p$  the observer can set up a local inertial coordinate system (such as normal coordinates) with coordinate radius of the order of  $B^{-\frac{1}{2}}$ . He can then choose a family  $\{f_i\}$  which satisfy equation (1.2) and which in the neighbourhood  $U$  are approximately positive frequency with respect to the time coordinate in  $U$ . For modes  $f_i$  whose characteristic frequency  $\omega$  is high compared to  $B^{\frac{1}{2}}$ , this leaves an indeterminacy between  $f_i$  and its complex conjugate  $\bar{f}_i$  of the order of the exponential of some multiple of  $-\omega B^{-\frac{1}{2}}$ . The indeterminacy between the annihilation operator  $a_i$  and the creation operator  $a_i^\dagger$  for the

mode is thus exponentially small. However, the ambiguity between the  $a_i$  and the  $a_i^\dagger$  is virtually complete for modes for which  $\omega < B^{\frac{1}{2}}$ . This ambiguity introduces an uncertainty of  $\pm \frac{1}{2}$  in the number operator  $a_i^\dagger a_i$  for the mode. The density of modes per unit volume in the frequency interval  $\omega$  to  $\omega + d\omega$  is of the order of  $\omega^2 d\omega$  for  $\omega$  greater than the rest mass  $m$  of the field in question. Thus the uncertainty in the local energy density caused by the ambiguity in defining modes of wavelength longer than the local radius of curvature  $B^{-\frac{1}{2}}$ , is of order  $B^2$  in units in which  $G = c = \hbar = 1$ . Because the ambiguity is exponentially small for wavelengths short compared to the radius of curvature  $B^{-\frac{1}{2}}$ , the total uncertainty in the local energy density is of order  $B^2$ . This uncertainty can be thought of as corresponding to the local energy density of particles created by the gravitational field. The uncertainty in the curvature produced via the Einstein equations by this uncertainty in the energy density is small compared to the total curvature of space-time provided that  $B$  is small compared to one, i.e. the radius of curvature  $B^{-\frac{1}{2}}$  is large compared to the Planck length  $10^{-33}$  cm. One would therefore expect that the scheme of treating the matter fields quantum mechanically on a classical curved space-time background would be a good approximation, except in regions where the curvature was comparable to the Planck value of  $10^{66}$  cm $^{-2}$ . From the classical singularity theorems [3–6], one would expect such high curvatures to occur in collapsing stars and, in the past, at the beginning of the present expansion phase of the universe. In the former case, one would expect the regions of high curvature to be hidden from us by an event horizon [7]. Thus, as far as we are concerned, the classical geometry–quantum matter treatment should be valid apart from the first  $10^{-43}$  s of the universe. The view is sometimes expressed that this treatment will break down when the radius of curvature is comparable to the Compton wavelength  $\sim 10^{-13}$  cm of an elementary particle such as a proton. However the Compton wavelength of a zero rest mass particle such as a photon or a neutrino is infinite, but we do not have any problem in dealing with electromagnetic or neutrino radiation in curved space-time. All that happens when the radius of curvature of space-time is smaller than the Compton wavelength of a given species of particle is that one gets an indeterminacy in the particle number or, in other words, particle creation. However, as was shown above, the energy density of the created particles is small locally compared to the curvature which created them.

Even though the effects of particle creation may be negligible locally, I shall show in this paper that they can add up to have a significant influence on black holes over the lifetime of the universe  $\sim 10^{17}$  s or  $10^{60}$  units of Planck time. It seems that the gravitational field of a black hole will create particles and emit them to infinity at just the rate that one would expect if the black hole were an ordinary body with a temperature in geometric units of  $\kappa/2\pi$ , where  $\kappa$  is the “surface gravity” of the black hole [8]. In ordinary units this temperature is of the order of  $10^{26} M^{-1} \text{ }^{\circ}\text{K}$ , where  $M$  is the mass, in grams of the black hole. For a black hole of solar mass ( $10^{33}$  g) this temperature is much lower than the  $3 \text{ }^{\circ}\text{K}$  temperature of the cosmic microwave background. Thus black holes of this size would be absorbing radiation faster than they emitted it and would be increasing in mass. However, in addition to black holes formed by stellar collapse, there might also be much smaller black holes which were formed by density fluctua-

tions in the early universe [9, 10]. These small black holes, being at a higher temperature, would radiate more than they absorbed. They would therefore presumably decrease in mass. As they got smaller, they would get hotter and so would radiate faster. As the temperature rose, it would exceed the rest mass of particles such as the electron and the muon and the black hole would begin to emit them also. When the temperature got up to about  $10^{12}$  °K or when the mass got down to about  $10^{14}$  g the number of different species of particles being emitted might be so great [11] that the black hole radiated away all its remaining rest mass on a strong interaction time scale of the order of  $10^{-23}$  s. This would produce an explosion with an energy of  $10^{35}$  ergs. Even if the number of species of particle emitted did not increase very much, the black hole would radiate away all its mass in the order of  $10^{-28} M^3$  s. In the last tenth of a second the energy released would be of the order of  $10^{30}$  ergs.

As the mass of the black hole decreased, the area of the event horizon would have to go down, thus violating the law that, classically, the area cannot decrease [7, 12]. This violation must, presumably, be caused by a flux of negative energy across the event horizon which balances the positive energy flux emitted to infinity. One might picture this negative energy flux in the following way. Just outside the event horizon there will be virtual pairs of particles, one with negative energy and one with positive energy. The negative particle is in a region which is classically forbidden but it can tunnel through the event horizon to the region inside the black hole where the Killing vector which represents time translations is spacelike. In this region the particle can exist as a real particle with a timelike momentum vector even though its energy relative to infinity as measured by the time translation Killing vector is negative. The other particle of the pair, having a positive energy, can escape to infinity where it constitutes a part of the thermal emission described above. The probability of the negative energy particle tunnelling through the horizon is governed by the surface gravity  $\kappa$  since this quantity measures the gradient of the magnitude of the Killing vector or, in other words, how fast the Killing vector is becoming spacelike. Instead of thinking of negative energy particles tunnelling through the horizon in the positive sense of time one could regard them as positive energy particles crossing the horizon on past-directed world-lines and then being scattered on to future-directed world-lines by the gravitational field. It should be emphasized that these pictures of the mechanism responsible for the thermal emission and area decrease are heuristic only and should not be taken too literally. It should not be thought unreasonable that a black hole, which is an excited state of the gravitational field, should decay quantum mechanically and that, because of quantum fluctuation of the metric, energy should be able to tunnel out of the potential well of a black hole. This particle creation is directly analogous to that caused by a deep potential well in flat space-time [18]. The real justification of the thermal emission is the mathematical derivation given in Section (2) for the case of an uncharged non-rotating black hole. The effects of angular momentum and charge are considered in Section (3). In Section (4) it is shown that any renormalization of the energy-momentum tensor with suitable properties must give a negative energy flow down the black hole and consequent decrease in the area of the event horizon. This negative energy flow is non-observable locally.

The decrease in area of the event horizon is caused by a violation of the weak energy condition [5–7, 12] which arises from the indeterminacy of particle number and energy density in a curved space-time. However, as was shown above, this indeterminacy is small, being of the order of  $B^2$  where  $B$  is the magnitude of the curvature tensor. Thus it can have a diverging effect on a null surface like the event horizon which has very small convergence or divergence but it can not untrap a strongly converging trapped surface until  $B$  becomes of the order of one. Therefore one would not expect the negative energy density to cause a breakdown of the classical singularity theorems until the radius of curvature of space-time became  $10^{-33}$  cm.

Perhaps the strongest reason for believing that black holes can create and emit particles at a steady rate is that the predicted rate is just that of the thermal emission of a body with the temperature  $\kappa/2\pi$ . There are independent, thermodynamic, grounds for regarding some multiple of the surface gravity as having a close relation to temperature. There is an obvious analogy with the second law of thermodynamics in the law that, classically, the area of the event horizon can never decrease and that when two black holes collide and merge together, the area of the final event horizon is greater than the sum of the areas of the two original horizons [7, 12]. There is also an analogy to the first law of thermodynamics in the result that two neighbouring black hole equilibrium states are related by [8]

$$dM = \frac{\kappa}{8\pi} dA + \Omega dJ$$

where  $M$ ,  $\Omega$ , and  $J$  are respectively the mass, angular velocity and angular momentum of the black hole and  $A$  is the area of the event horizon. Comparing this to

$$dU = TdS + pdV$$

one sees that if some multiple of  $A$  is regarded as being analogous to entropy, then some multiple of  $\kappa$  is analogous to temperature. The surface gravity is also analogous to temperature in that it is constant over the event horizon in equilibrium. Beckenstein [19] suggested that  $A$  and  $\kappa$  were not merely analogous to entropy and temperature respectively but that, in some sense, they actually were the entropy and temperature of the black hole. Although the ordinary second law of thermodynamics is transcended in that entropy can be lost down black holes, the flow of entropy across the event horizon would always cause some increase in the area of the horizon. Beckenstein therefore suggested [20] a Generalized Second Law: Entropy + some multiple (unspecified) of  $A$  never decreases. However he did not suggest that a black hole could emit particles as well as absorb them. Without such emission the Generalized Second Law would be violated by for example, a black hole immersed in black body radiation at a lower temperature than that of the black hole. On the other hand, if one accepts that black holes do emit particles at a steady rate, the identification of  $\kappa/2\pi$  with temperature and  $\frac{1}{4}A$  with entropy is established and a Generalized Second Law confirmed.

## 2. Gravitational Collapse

It is now generally believed that, according to classical theory, a gravitational collapse will produce a black hole which will settle down rapidly to a stationary axisymmetric equilibrium state characterized by its mass, angular momentum and electric charge [7, 13]. The Kerr-Newman solution represent one such family of black hole equilibrium states and it seems unlikely that there are any others. It has therefore become a common practice to ignore the collapse phase and to represent a black hole simply by one of these solutions. Because these solutions are stationary there will not be any mixing of positive and negative frequencies and so one would not expect to obtain any particle creation. However there is a classical phenomenon called superradiance [14–17] in which waves incident in certain modes on a rotating or charged black hole are scattered with increased amplitude [see Section (3)]. On a particle description this amplification must correspond to an increase in the number of particles and therefore to stimulated emission of particles. One would therefore expect on general grounds that there would also be a steady rate of spontaneous emission in these superradiant modes which would tend to carry away the angular momentum or charge of the black hole [16]. To understand how the particle creation can arise from mixing of positive and negative frequencies, it is essential to consider not only the quasi-stationary final state of the black hole but also the time-dependent formation phase. One would hope that, in the spirit of the “no hair” theorems, the rate of emission would not depend on details of the collapse process except through the mass, angular momentum and charge of the resulting black hole. I shall show that this is indeed the case but that, in addition to the emission in the superradiant modes, there is a steady rate of emission in all modes at the rate one would expect if the black hole were an ordinary body with temperature  $\kappa/2\pi$ .

I shall consider first of all the simplest case of a non-rotating uncharged black hole. The final stationary state for such a black hole is represented by the Schwarzschild solution with metric

$$ds^2 = - \left(1 - \frac{2M}{r}\right) dt^2 + \left(1 - \frac{2M}{r}\right)^{-1} dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2). \quad (2.1)$$

As is now well known, the apparent singularities at  $r=2M$  are fictitious, arising merely from a bad choice of coordinates. The global structure of the analytically extended Schwarzschild solution can be described in a simple manner by a Penrose diagram of the  $r-t$  plane (Fig. 1) [6, 13]. In this diagram null geodesics in the  $r-t$  plane are at  $\pm 45^\circ$  to the vertical. Each point of the diagram represents a 2-sphere of area  $4\pi r^2$ . A conformal transformation has been applied to bring infinity to a finite distance: infinity is represented by the two diagonal lines (really null surfaces) labelled  $\mathcal{I}^+$  and  $\mathcal{I}^-$ , and the points  $I^+$ ,  $I^-$ , and  $I^0$ . The two horizontal lines  $r=0$  are curvature singularities and the two diagonal lines  $r=2M$  (really null surfaces) are the future and past event horizons which divide the solution up into regions from which one cannot escape to  $\mathcal{I}^+$  and  $\mathcal{I}^-$ . On the left of the diagram there is another infinity and asymptotically flat region.

Most of the Penrose diagram is not in fact relevant to a black hole formed by gravitational collapse since the metric is that of the Schwarzschild solution

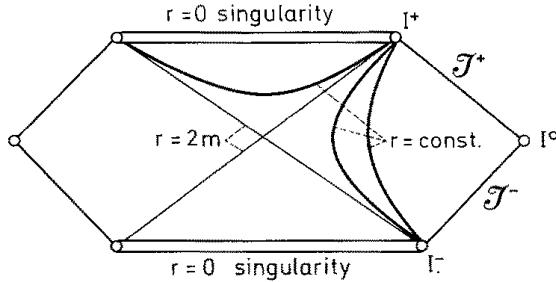


Fig. 1. The Penrose diagram for the analytically extended Schwarzschild solution

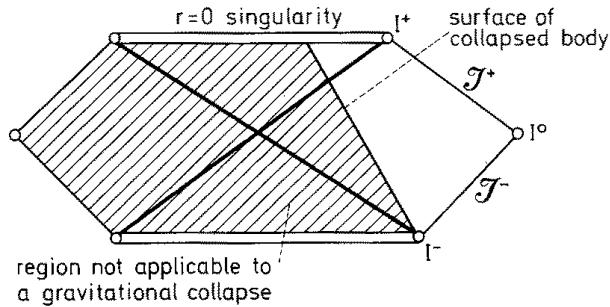


Fig. 2. Only the region of the Schwarzschild solution outside the collapsing body is relevant for a black hole formed by gravitational collapse. Inside the body the solution is completely different

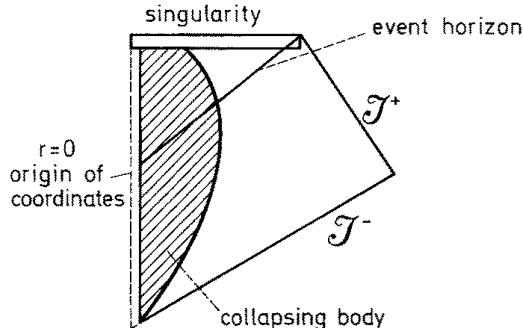


Fig. 3. The Penrose diagram of a spherically symmetric collapsing body producing a black hole. The vertical dotted line on the left represents the non-singular centre of the body

only in the region outside the collapsing matter and only in the asymptotic future. In the case of exactly spherical collapse, which I shall consider for simplicity, the metric is exactly the Schwarzschild metric everywhere outside the surface of the collapsing object which is represented by a timelike geodesic in the Penrose diagram (Fig. 2). Inside the object the metric is completely different, the past event horizon, the past  $r=0$  singularity and the other asymptotically flat region do not exist and are replaced by a time-like curve representing the origin of polar coordinates. The appropriate Penrose diagram is shown in Fig. 3 where the conformal freedom has been used to make the origin of polar coordinates into a vertical line.

In this space-time consider (again for simplicity) a massless Hermitian scalar field operator  $\phi$  obeying the wave equation

$$\phi_{;ab}g^{ab}=0. \quad (2.2)$$

(The results obtained would be the same if one used the conformally invariant wave equation:

$$\phi_{;ab}g^{ab} + \frac{1}{6}R\phi = 0.)$$

The operator  $\phi$  can be expressed as

$$\phi = \sum_i \{f_i \mathbf{a}_i + \bar{f}_i \mathbf{a}_i^\dagger\}. \quad (2.3)$$

The solutions  $\{f_i\}$  of the wave equation  $f_{i;ab}g^{ab}=0$  can be chosen so that on past null infinity  $\mathcal{I}^-$  they form a complete family satisfying the orthonormality conditions (1.2) where the surface  $S$  is  $\mathcal{I}^-$  and so that they contain only positive frequencies with respect to the canonical affine parameter on  $\mathcal{I}^-$ . (This last condition of positive frequency can be uniquely defined despite the existence of “supertranslations” in the Bondi-Metzner-Sachs asymptotic symmetry group [21, 22].) The operators  $\mathbf{a}_i$  and  $\mathbf{a}_i^\dagger$  have the natural interpretation as the annihilation and creation operators for ingoing particles i.e. for particles at past null infinity  $\mathcal{I}^-$ . Because massless fields are completely determined by their data on  $\mathcal{I}^-$ , the operator  $\phi$  can be expressed in the form (2.3) everywhere. In the region outside the event horizon one can also determine massless fields by their data on the event horizon and on future null infinity  $\mathcal{I}^+$ . Thus one can also express  $\phi$  in the form

$$\phi = \sum_i \{p_i \mathbf{b}_i + \bar{p}_i \mathbf{b}_i^\dagger + q_i \mathbf{c}_i + \bar{q}_i \mathbf{c}_i^\dagger\}. \quad (2.4)$$

Here the  $\{p_i\}$  are solutions of the wave equation which are purely outgoing, i.e. they have zero Cauchy data on the event horizon and the  $\{q_i\}$  are solutions which contain no outgoing component, i.e. they have zero Cauchy data on  $\mathcal{I}^+$ . The  $\{p_i\}$  and  $\{q_i\}$  are required to be complete families satisfying the orthonormality conditions (1.2) where the surface  $S$  is taken to be  $\mathcal{I}^+$  and the event horizon respectively. In addition the  $\{p_i\}$  are required to contain only positive frequencies with respect to the canonical affine parameter along the null geodesic generators of  $\mathcal{I}^+$ . With the positive frequency condition on  $\{p_i\}$ , the operators  $\{\mathbf{b}_i\}$  and  $\{\mathbf{b}_i^\dagger\}$  can be interpreted as the annihilation and creation operators for outgoing particles, i.e. for particles on  $\mathcal{I}^+$ . It is not clear whether one should impose some positive frequency condition on the  $\{q_i\}$  and if so with respect to what. The choice of the  $\{q_i\}$  does not affect the calculation of the emission of particles to  $\mathcal{I}^+$ . I shall return to the question in Section (4).

Because massless fields are completely determined by their data on  $\mathcal{I}^-$  one can express  $\{p_i\}$  and  $\{q_i\}$  as linear combinations of the  $\{f_i\}$  and  $\{\bar{f}_i\}$ :

$$p_i = \sum_j (\alpha_{ij} f_j + \beta_{ij} \bar{f}_j), \quad (2.5)$$

$$q_i = \sum_j (\gamma_{ij} f_j + \eta_{ij} \bar{f}_j). \quad (2.6)$$

These relations lead to corresponding relations between the operators

$$\mathbf{b}_i = \sum_j (\bar{\alpha}_{ij} \mathbf{a}_j - \bar{\beta}_{ij} \mathbf{a}_j^\dagger), \quad (2.7)$$

$$\mathbf{c}_i = \sum_j (\bar{\gamma}_{ij} \mathbf{a}_j - \bar{\eta}_{ij} \mathbf{a}_j^\dagger). \quad (2.8)$$

The initial vacuum state  $|0\rangle$ , the state containing no incoming particles, i.e. no particles on  $\mathcal{I}^-$ , is defined by

$$\alpha_i |0\rangle = 0 \quad \text{for all } i. \quad (2.9)$$

However, because the coefficients  $\beta_{ij}$  will not be zero in general, the initial vacuum state will not appear to be a vacuum state to an observer at  $\mathcal{I}^+$ . Instead he will find that the expectation value of the number operator for the  $i$ th outgoing mode is

$$\langle 0_- | b_i^\dagger b_i | 0_- \rangle = \sum_j |\beta_{ij}|^2. \quad (2.10)$$

Thus in order to determine the number of particles created by the gravitational field and emitted to infinity one simply has to calculate the coefficients  $\beta_{ij}$ . One would expect this calculation to be very messy and to depend on the detailed nature of the gravitational collapse. However, as I shall show, one can derive an asymptotic form for the  $\beta_{ij}$  which depends only on the surface gravity of the resulting black hole. There will be a certain finite amount of particle creation which depends on the details of the collapse. These particles will disperse and at late retarded times on  $\mathcal{I}^+$  there will be a steady flux of particles determined by the asymptotic form of  $\beta_{ij}$ .

In order to calculate this asymptotic form it is more convenient to decompose the ingoing and outgoing solutions of the wave equation into their Fourier components with respect to advanced or retarded time and use the continuum normalization. The finite normalization solutions can then be recovered by adding Fourier components to form wave packets. Because the space-time is spherically symmetric, one can also decompose the incoming and outgoing solutions into spherical harmonics. Thus, in the region outside the collapsing body, one can write the incoming and outgoing solutions as

$$f_{\omega' lm} = (2\pi)^{-\frac{1}{2}} r^{-1} (\omega')^{-\frac{1}{2}} F_{\omega'}(r) e^{i\omega' v} Y_{lm}(\theta, \phi), \quad (2.11)$$

$$p_{\omega lm} = (2\pi)^{-\frac{1}{2}} r^{-1} \omega^{-\frac{1}{2}} P_\omega(r) e^{i\omega u} Y_{lm}(\theta, \phi), \quad (2.12)$$

where  $v$  and  $u$  are the usual advanced and retarded coordinates defined by

$$v = t + r + 2M \log \left| \frac{r}{2M} - 1 \right|, \quad (2.13)$$

$$u = t - r - 2M \log \left| \frac{r}{2M} - 1 \right|. \quad (2.14)$$

Each solution  $p_{\omega lm}$  can be expressed as an integral with respect to  $\omega'$  over solutions  $f_{\omega' lm}$  and  $\bar{f}_{\omega' lm}$  with the same values of  $l$  and  $|m|$  (from now on I shall drop the suffices  $l, m$ ):

$$p_\omega = \int_0^\infty (\alpha_{\omega\omega'} f_{\omega'} + \beta_{\omega\omega'} \bar{f}_{\omega'}) d\omega'. \quad (2.15)$$

To calculate the coefficients  $\alpha_{\omega\omega'}$  and  $\beta_{\omega\omega'}$ , consider a solution  $p_\omega$  propagating backwards from  $\mathcal{I}^+$  with zero Cauchy data on the event horizon. A part  $p_\omega^{(1)}$  of the solution  $p_\omega$  will be scattered by the static Schwarzschild field outside the collapsing body and will end up on  $\mathcal{I}^-$  with the same frequency  $\omega$ . This will give a  $\delta(\omega' - \omega)$  term in  $\alpha_{\omega\omega'}$ . The remainder  $p_\omega^{(2)}$  of  $p_\omega$  will enter the collapsing body

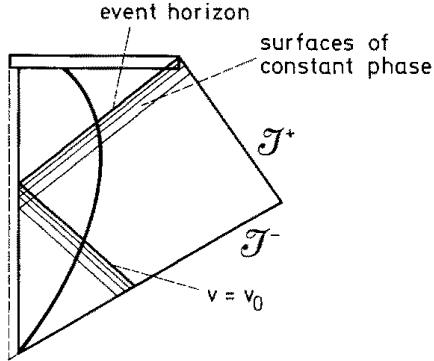


Fig. 4. The solution  $p_\omega$  of the wave equation has an infinite number of cycles near the event horizon and near the surface  $v=v_0$

where it will be partly scattered and partly reflected through the centre, eventually emerging to  $\mathcal{I}^-$ . It is this part  $p_\omega^{(2)}$  which produces the interesting effects. Because the retarded time coordinate  $u$  goes to infinity on the event horizon, the surfaces of constant phase of the solution  $p_\omega$  will pile up near the event horizon (Fig. 4). To an observer on the collapsing body the wave would seem to have a very large blue-shift. Because its effective frequency was very high, the wave would propagate by geometric optics through the centre of the body and out on  $\mathcal{I}^-$ . On  $\mathcal{I}^-$   $p_\omega^{(2)}$  would have an infinite number of cycles just before the advanced time  $v=v_0$  where  $v_0$  is the latest time that a null geodesic could leave  $\mathcal{I}^-$ , pass through the centre of the body and escape to  $\mathcal{I}^+$  before being trapped by the event horizon. One can estimate the form of  $p_\omega^{(2)}$  on  $\mathcal{I}^-$  near  $v=v_0$  in the following way. Let  $x$  be a point on the event horizon outside the matter and let  $l^a$  be a null vector tangent to the horizon. Let  $n^a$  be the future-directed null vector at  $x$  which is directed radially inwards and normalized so that  $l^a n_a = -1$ . The vector  $-\epsilon n^a$  ( $\epsilon$  small and positive) will connect the point  $x$  on the event horizon with a nearby null surface of constant retarded time  $u$  and therefore with a surface of constant phase of the solution  $p_\omega^{(2)}$ . If the vectors  $l^a$  and  $n^a$  are parallelly transported along the null geodesic  $\gamma$  through  $x$  which generates the horizon, the vector  $-\epsilon n^a$  will always connect the event horizon with the same surface of constant phase of  $p_\omega^{(2)}$ . To see what the relation between  $\epsilon$  and the phase of  $p_\omega^{(2)}$  is, imagine in Fig. 2 that the collapsing body did not exist but one analytically continued the empty space Schwarzschild solution back to cover the whole Penrose diagram. One could then transport the pair  $(l^a, n^a)$  back along to the point where future and past event horizons intersected. The vector  $-\epsilon n^a$  would then lie along the past event horizon. Let  $\lambda$  be the affine parameter along the past event horizon which is such that at the point of intersection of the two horizons,  $\lambda=0$  and  $\frac{dx^a}{d\lambda}=n^a$ . The affine parameter  $\lambda$  is related to the retarded time  $u$  on the past horizon by

$$\lambda = -C e^{-\kappa u} \quad (2.16)$$

where  $C$  is constant and  $\kappa$  is the surface gravity of the black hole defined by  $K_{;b}^a K^b = -\kappa K^a$  on the horizon where  $K^a$  is the time translation Killing vector.

For a Schwarzschild black hole  $\kappa = \frac{1}{4M}$ . It follows from this that the vector  $-\varepsilon n^a$  connects the future event horizon with the surface of constant phase  $-\frac{\omega}{\kappa}(\log \varepsilon - \log C)$  of the solution  $p_\omega^{(2)}$ . This result will also hold in the real space-time (including the collapsing body) in the region outside the body. Near the event horizon the solution  $p_\omega^{(2)}$  will obey the geometric optics approximation as it passes through the body because its effective frequency will be very high. This means that if one extends the null geodesic  $\gamma$  back past the end-point of the event horizon and out onto  $\mathcal{I}^-$  at  $v=v_0$  and parallelly transports  $n^a$  along  $\gamma$ , the vector  $-\varepsilon n^a$  will still connect  $\gamma$  to a surface of constant phase of the solution  $p_\omega^{(2)}$ . On  $\mathcal{I}^- n^a$  will be parallel to the Killing vector  $K^a$  which is tangent to the null geodesic generators of  $\mathcal{I}^-$ :

$$n^a = DK^a.$$

Thus on  $\mathcal{I}^-$  for  $v_0 - v$  small and positive, the phase of the solution will be

$$-\frac{\omega}{\kappa}(\log(v_0 - v) - \log D - \log C). \quad (2.17)$$

Thus on  $\mathcal{I}^- p_\omega^{(2)}$  will be zero for  $v > v_0$  and for  $v < v_0$

$$p_\omega^{(2)} \sim (2\pi)^{-\frac{1}{2}} \omega^{-\frac{1}{2}} r^{-1} P_\omega^- \exp\left(-i\frac{\omega}{\kappa} \left(\log\left(\frac{v_0 - v}{CD}\right)\right)\right) \quad (2.18)$$

where  $P_\omega^- \equiv P_\omega(2M)$  is the value of the radial function for  $P_\omega$  on the past event horizon in the analytically continued Schwarzschild solution. The expression (2.18) for  $p_\omega^{(2)}$  is valid only for  $v_0 - v$  small and positive. At earlier advanced times the amplitude will be different and the frequency measured with respect to  $v$ , will approach the original frequency  $\omega$ .

By Fourier transforming  $p_\omega^{(2)}$  one can evaluate its contributions to  $\alpha_{\omega\omega'}$  and  $\beta_{\omega\omega'}$ . For large values of  $\omega'$  these will be determined by the asymptotic form (2.18). Thus for large  $\omega'$

$$\alpha_{\omega\omega'}^{(2)} \approx (2\pi)^{-1} P_\omega^-(CD)^{\frac{i\omega}{\kappa}} \exp(i(\omega - \omega')v_0) \left(\frac{\omega'}{\omega}\right)^{\frac{1}{2}} \Gamma\left(1 - \frac{i\omega}{\kappa}\right) (-i\omega')^{-1 + \frac{i\omega}{\kappa}}, \quad (2.19)$$

$$\beta_{\omega\omega'}^{(2)} \approx -i\alpha_{\omega(-\omega')}^{(2)}. \quad (2.20)$$

The solution  $p_\omega^{(2)}$  is zero on  $\mathcal{I}^-$  for large values of  $v$ . This means that its Fourier transform is analytic in the upper half  $\omega'$  plane and that  $p_\omega^{(2)}$  will be correctly represented by a Fourier integral in which the contour has been displaced into the upper half  $\omega'$  plane. The Fourier transform of  $p_\omega^{(2)}$  contains a factor  $(-i\omega')^{-1 + \frac{i\omega}{\kappa}}$  which has a logarithmic singularity at  $\omega' = 0$ . To obtain  $\beta_{\omega\omega'}^{(2)}$  from  $\alpha_{\omega\omega'}^{(2)}$  by (2.20) one has to analytically continue  $\alpha_{\omega\omega'}^{(2)}$  anticlockwise round this singularity. This means that

$$|\alpha_{\omega\omega'}^{(2)}| = \exp\left(\frac{\pi\omega}{\kappa}\right) |\beta_{\omega\omega'}^{(2)}|. \quad (2.21)$$

Actually, the fact that  $p_{\omega}^{(2)}$  is not given by (2.18) at early advanced times means that the singularity in  $\alpha_{\omega\omega'}$  occurs at  $\omega'=\omega$  and not at  $\omega'=0$ . However the relation (2.21) is still valid for large  $\omega'$ .

The expectation value of the total number of created particles at  $\mathcal{I}^+$  in the frequency range  $\omega$  to  $\omega+d\omega$  is  $d\omega \int_0^\infty |\beta_{\omega\omega'}|^2 d\omega'$ . Because  $|\beta_{\omega\omega'}|$  goes like  $(\omega')^{-\frac{1}{2}}$  at large  $\omega'$  this integral diverges. This infinite total number of created particles corresponds to a finite steady rate of emission continuing for an infinite time as can be seen by building up a complete orthonormal family of wave packets from the Fourier components  $p_{\omega}$ . Let

$$p_{jn} = \varepsilon^{-\frac{1}{2}} \int_{j\varepsilon}^{(j+1)\varepsilon} e^{-2\pi i n \varepsilon^{-1} \omega} p_{\omega} d\omega \quad (2.22)$$

where  $j$  and  $n$  are integers,  $j \geq 0$ ,  $\varepsilon > 0$ . For  $\varepsilon$  small these wave packets will have frequency  $j\varepsilon$  and will be peaked around retarded time  $u = 2\pi n \varepsilon^{-1}$  with width  $\varepsilon^{-1}$ . One can expand  $\{p_{jn}\}$  in terms of the  $\{f_{\omega}\}$

$$p_{jn} = \int_0^\infty (\alpha_{jn\omega'} f_{\omega'} + \beta_{jn\omega'} \bar{f}_{\omega'}) d\omega' \quad (2.23)$$

where

$$\alpha_{jn\omega'} = \varepsilon^{-\frac{1}{2}} \int_{j\varepsilon}^{(j+1)\varepsilon} e^{-2\pi i n \varepsilon^{-1} \omega} \alpha_{\omega\omega'} d\omega \quad \text{etc.} \quad (2.24)$$

For  $j \gg \varepsilon$ ,  $n \gg \varepsilon$

$$\begin{aligned} |\alpha_{jn\omega'}| &= \left| (2\pi)^{-1} P_{\omega}^- \omega^{-\frac{1}{2}} \Gamma\left(1 - \frac{i\omega}{\kappa}\right) \varepsilon^{-\frac{1}{2}} (\omega')^{-\frac{1}{2}} \right. \\ &\quad \cdot \left. \int_{j\varepsilon}^{(j+1)\varepsilon} \exp i\omega'' (-2\pi n \varepsilon^{-1} + \kappa^{-1} \log \omega') d\omega'' \right| \\ &= \left| \pi^{-1} P_{\omega}^- \omega^{-\frac{1}{2}} \Gamma\left(1 - \frac{i\omega}{\kappa}\right) \varepsilon^{-\frac{1}{2}} (\omega')^{-\frac{1}{2}} z^{-1} \sin \frac{1}{2} \varepsilon z \right| \end{aligned} \quad (2.25)$$

where  $\omega = j\varepsilon$  and  $z = \kappa^{-1} \log \omega' - 2\pi n \varepsilon^{-1}$ . For wave-packets which reach  $\mathcal{I}^+$  at late retarded times, i.e. those with large values of  $n$ , the main contribution to  $\alpha_{jn\omega'}$  and  $\beta_{jn\omega'}$  come from very high frequencies  $\omega'$  of the order of  $\exp(2\pi n \kappa \varepsilon^{-1})$ . This means that these coefficients are governed only by the asymptotic forms (2.19, 2.20) for high  $\omega'$  which are independent of the details of the collapse.

The expectation value of the number of particles created and emitted to infinity  $\mathcal{I}^+$  in the wave-packet mode  $p_{jn}$  is

$$\int_0^\infty |\beta_{jn\omega'}|^2 d\omega'. \quad (2.26)$$

One can evaluate this as follows. Consider the wave-packet  $p_{jn}$  propagating backwards from  $\mathcal{I}^+$ . A fraction  $1 - \Gamma_{jn}$  of the wave-packet will be scattered by the static Schwarzschild field and a fraction  $\Gamma_{jn}$  will enter the collapsing body.

$$\Gamma_{jn} = \int_0^\infty (|\alpha_{jn\omega'}^{(2)}|^2 - |\beta_{jn\omega'}^{(2)}|^2) d\omega' \quad (2.27)$$

where  $\alpha_{jn\omega'}^{(2)}$  and  $\beta_{jn\omega'}^{(2)}$  are calculated using (2.19, 2.20) from the part  $p_{jn}^{(2)}$  of the wave-packet which enters the star. The minus sign in front of the second term on the right of (2.27) occurs because the negative frequency components of  $p_{jn}^{(2)}$  make a negative contribution to the flux into the collapsing body. By (2.21)

$$|\alpha_{jn\omega'}^{(2)}| = \exp(\pi \omega \kappa^{-1}) |\beta_{jn\omega'}^{(2)}|. \quad (2.28)$$

Thus the total number of particles created in the mode  $p_{jn}$  is

$$\Gamma_{jn}(\exp(2\pi\omega\kappa^{-1}) - 1)^{-1}. \quad (2.29)$$

But for wave-packets at late retarded times, the fraction  $\Gamma_{jn}$  which enters the collapsing body is almost the same as the fraction of the wave-packet that would have crossed the past event horizon had the collapsing body not been there but the exterior Schwarzschild solution had been analytically continued. Thus this factor  $\Gamma_{jn}$  is also the same as the fraction of a similar wave-packet coming from  $\mathcal{I}^-$  which would have crossed the future event horizon and have been absorbed by the black hole. The relation between emission and absorption cross-section is therefore exactly that for a body with a temperature, in geometric units, of  $\kappa/2\pi$ .

Similar results hold for the electromagnetic and linearised gravitational fields. The fields produced on  $\mathcal{I}^-$  by positive frequency waves from  $\mathcal{I}^+$  have the same asymptotic form as (2.18) but with an extra blue shift factor in the amplitude. This extra factor cancels out in the definition of the scalar product so that the asymptotic forms of the coefficients  $\alpha$  and  $\beta$  are the same as in the Eqs. (2.19) and (2.20). Thus one would expect the black hole also to radiate photons and gravitons thermally. For massless fermions such as neutrinos one again gets similar results except that the negative frequency components given by the coefficients  $\beta$  now make a positive contribution to the probability flux into the collapsing body. This means that the term  $|\beta|^2$  in (2.27) now has the opposite sign. From this it follows that the number of particles emitted in any outgoing wave packet mode is  $(\exp(2\pi\omega\kappa^{-1}) + 1)^{-1}$  times the fraction of that wave packet that would have been absorbed by the black hole had it been incident from  $\mathcal{I}^-$ . This is again exactly what one would expect for thermal emission of particles obeying Fermi-Dirac statistics.

Fields of non-zero rest mass do not reach  $\mathcal{I}^-$  and  $\mathcal{I}^+$ . One therefore has to describe ingoing and outgoing states for these fields in terms of some concept such as the projective infinity of Eardley and Sachs [23] and Schmidt [24]. However, if the initial and final states are asymptotically Schwarzschild or Kerr solutions, one can describe the ingoing and outgoing states in a simple manner by separation of variables and one can define positive frequencies with respect to the time translation Killing vectors of these initial and final asymptotic space-times. In the asymptotic future there will be no bound states: any particle will either fall through the event horizon or escape to infinity. Thus the unbound outgoing states and the event horizon states together form a complete basis for solutions of the wave equation in the region outside the event horizon. In the asymptotic past there could be bound states if the body that collapses had had a bounded radius for an infinite time. However one could equally well assume that the body had collapsed from an infinite radius in which case there would be no bound states. The possible existence of bound states in the past does not affect the rate of particle emission in the asymptotic future which will again be that of a body with temperature  $\kappa/2\pi$ . The only difference from the zero rest mass case is that the frequency  $\omega$  in the thermal factor  $(\exp(2\pi\omega\kappa^{-1}) \mp 1)^{-1}$  now includes the rest mass energy of the particle. Thus there will not be much emission of particles of rest mass  $m$  unless the temperature  $\kappa/2\pi$  is greater than  $m$ .

One can show that these results on thermal emission do not depend on spherical symmetry. Consider an asymmetric collapse which produced a black hole which settled to a non-rotating uncharged Schwarzschild solution (angular momentum and charge will be considered in the next section). The fact that the final state is asymptotically quasi-stationary means that there is a preferred Bondi coordinate system [25] on  $\mathcal{I}^+$  with respect to which one can decompose the Cauchy data for the outgoing states into positive frequencies and spherical harmonics. On  $\mathcal{I}^-$  there may or may not be a preferred coordinate system but if there is not one can pick an arbitrary Bondi coordinate system and decompose the Cauchy data for the ingoing states in a similar manner. Now consider one of the  $\mathcal{I}^+$  states  $p_{\omega lm}$  propagating backwards through this space-time into the collapsing body and out again onto  $\mathcal{I}^-$ . Take a null geodesic generator  $\gamma$  of the event horizon and extend it backwards beyond its past end-point to intersect  $\mathcal{I}^-$  at a point  $y$  on a null geodesic generator  $\lambda$  of  $\mathcal{I}^-$ . Choose a pair of null vectors  $(l^a, \hat{n}^a)$  at  $y$  with  $l^a$  tangent to  $\gamma$  and  $\hat{n}^a$  tangent to  $\lambda$ . Parallelly propagate  $l^a, \hat{n}^a$  along  $\gamma$  to a point  $x$  in the region of space-time where the metric is almost that of the final Schwarzschild solution. At  $x$   $\hat{n}^a$  will be some linear combination of  $l^a$  and the radial inward directed null vector  $n^a$ . This means that the vector  $-\varepsilon \hat{n}^a$  will connect  $x$  to a surface of phase  $-\omega/\kappa (\log \varepsilon - \log E)$  of the solution  $p_{\omega lm}$  where  $E$  is some constant. As before, by the geometric optics approximation, the vector  $-\varepsilon \hat{n}^a$  at  $y$  will connect  $y$  to a surface of phase  $-\omega/\kappa (\log \varepsilon - \log E)$  of  $p_{\omega lm}^{(2)}$  where  $p_{\omega lm}^{(2)}$  is the part of  $p_{\omega lm}$  which enters the collapsing body. Thus on the null geodesic generator  $\lambda$  of  $\mathcal{I}^-$ , the phase of  $p_{\omega lm}^{(2)}$  will be

$$-\frac{i\omega}{\kappa} (\log(v_0 - v) - \log H) \quad (2.30)$$

where  $v$  is an affine parameter on  $\lambda$  with value  $v_0$  at  $y$  and  $H$  is a constant. By the geometrical optics approximation, the value of  $p_{\omega lm}^{(2)}$  on  $\lambda$  will be

$$L \exp \left\{ -\frac{i\omega}{\kappa} [\log(v_0 - v) - \log H] \right\} \quad (2.31)$$

for  $v_0 - v$  small and positive and zero for  $v > v_0$  where  $L$  is a constant. On each null geodesic generator of  $\mathcal{I}^-$   $p_{\omega lm}^{(2)}$  will have the form (2.31) with different values of  $L$ ,  $v_0$ , and  $H$ . The lack of spherical symmetry during the collapse will cause  $p_{\omega lm}^{(2)}$  on  $\mathcal{I}^-$  to contain components of spherical harmonics with indices  $(l', m')$  different from  $(l, m)$ . This means that one now has to express  $p_{\omega lm}^{(2)}$  in the form

$$p_{\omega lm}^{(2)} = \sum_{l' m'} \int_0^\infty \{ \alpha_{\omega lm \omega' l' m'}^{(2)} f_{\omega' l' m'} + \beta_{\omega lm \omega' l' m'}^{(2)} \bar{f}_{\omega' l' m'} \} d\omega'. \quad (2.32)$$

Because of (2.31), the coefficients  $\alpha^{(2)}$  and  $\beta^{(2)}$  will have the same  $\omega'$  dependence as in (2.19) and (2.20). Thus one still has the same relation as (2.21):

$$|\alpha_{\omega lm \omega' l' m'}^{(2)}| = \exp(\pi \omega \kappa^{-1}) |\beta_{\omega lm \omega' l' m'}^{(2)}|. \quad (2.33)$$

As before, for each  $(l, m)$ , one can make up wave packets  $p_{jnlm}$ . The number of particles emitted in such a wave packet mode is

$$\sum_{l', m'} \int_0^\infty |\beta_{jnlm \omega' l' m'}| {}^2 d\omega'. \quad (2.34)$$

Similarly, the fraction  $\Gamma_{jnlm}$  of the wave packet that enters the collapsing body is

$$\Gamma_{jnlm} = \sum_{l', m'} \int_0^\infty \{ |\alpha_{jnlm\omega' l' m'}^{(2)}|^2 - |\beta_{jnlm\omega' l' m'}^{(2)}|^2 \} d\omega'. \quad (2.35)$$

Again,  $\Gamma_{jnlm}$  is equal to the fraction of a similar wave packet coming from  $\mathcal{I}^-$  that would have been absorbed by the black hole. Thus, using (2.33), one finds that the emission is just that of a body of temperature  $\kappa/2\pi$ : the emission at late retarded times depends only on the final quasi-stationary state of the black hole and not on the details of the gravitational collapse.

### 3. Angular Momentum and Charge

If the collapsing body was rotating or electrically charged, the resulting black hole would settle down to a stationary state which was described, not by the Schwarzschild solution, but by a charged Kerr solution characterised by the mass  $M$ , the angular momentum  $J$ , and the charge  $Q$ . As these solutions are stationary and axisymmetric, one can separate solutions of the wave equations in them into a factor  $e^{i\omega u}$  or  $e^{i\omega v}$  times  $e^{-im\phi}$  times a function of  $r$  and  $\theta$ . In the case of the scalar wave equation one can separate this last expression into a function of  $r$  times a function of  $\theta$  [26]. One can also completely separate any wave equation in the non-rotating charged case and Teukolsky [27] has obtained completely separable wave equations for neutrino, electromagnetic and linearised gravitational fields in the uncharged rotating case.

Consider a wave packet of a classical field of charge  $e$  with frequency  $\omega$  and axial quantum number  $m$  incident from infinity on a Kerr black hole. The change in mass  $dM$  of the black hole caused by the partial absorption of the wave packet will be related to the change in area, angular momentum and charge by the classical first law of black holes:

$$dM = \frac{\kappa}{8\pi} dA + \Omega dJ + \Phi dQ \quad (3.1)$$

where  $\Omega$  and  $\Phi$  are the angular frequency and electrostatic potential respectively of the black hole [13]. The fluxes of energy, angular momentum and charge in the wave packet will be in the ratio  $\omega:m:e$ . Thus the changes in the mass, angular momentum and charge of the black hole will also be in this ratio. Therefore

$$dM(1 - \Omega m \omega^{-1} - e \Phi \omega^{-1}) = \frac{\kappa}{8\pi} dA. \quad (3.2)$$

A wave packet of a classical Boson field will obey the weak energy condition: the local energy density for any observer is non-negative. It follows from this [7, 12] that the change in area  $dA$  induced by the wave-packet will be non-negative. Thus if

$$\omega < m\Omega + e\Phi \quad (3.3)$$

the change in mass  $dM$  of the black hole must be negative. In other words, the black hole will lose energy to the wave packet which will therefore be scattered with the same frequency but increased amplitude. This is the phenomenon known as “superradiance”.

For classical fields of half-integer spin, detailed calculations [28] show that there is no superradiance. The reason for this is that the scalar product for half-integer spin fields is positive definite unlike that for integer spins. This means that the probability flux across the event horizon is positive and therefore, by conservation of probability, the probability flux in the scattered wave packet must be less than that in the incident wave packet. The reason that the above argument based on the first law breaks down is that the energy-momentum tensor for a classical half-integer spin field does not obey the weak energy condition. On a quantum, particle level one can understand the absence of superradiance for fermion fields as a consequence of the fact that the Exclusion Principle does not allow more than one particle in each outgoing wave packet mode and therefore does not allow the scattered wave-packet to be stronger than the incident wave-packet.

Passing now to the quantum theory, consider first the case of an unchanged, rotating black hole. One can as before pick an arbitrary Bondi coordinate frame on  $\mathcal{I}^-$  and decompose the operator  $\phi$  in terms of a family  $\{f_{\omega lm}\}$  of incoming solutions where the indices  $\omega$ ,  $l$ , and  $m$  refer to the advanced time and angular dependence of  $f$  on  $\mathcal{I}^-$  in the given coordinate system. On  $\mathcal{I}^+$  the final quasi-stationary state of the black hole defines a preferred Bondi coordinate system using which one can define a family  $\{p_{\omega lm}\}$  of outgoing solutions. The index  $l$  in this case labels the spheroidal harmonics in terms of which the wave equation is separable. One proceeds as before to calculate the asymptotic form of  $p_{\omega lm}^{(2)}$  on  $\mathcal{I}^-$ . The only difference is that because the horizon is rotating with angular velocity  $\Omega$  with respect to  $\mathcal{I}^+$ , the effective frequency near a generator of the event horizon is not  $\omega$  but  $\omega - m\Omega$ . This means that the number of particles emitted in the wave-packet mode  $p_{jnlm}$  is

$$\{\exp(2\pi\kappa^{-1}(\omega - m\Omega)) \mp 1\}^{-1} \Gamma_{jnlm}. \quad (3.4)$$

The effect of this is to cause the rate of emission of particles with positive angular momentum  $m$  to be higher than that of particles with the same frequency  $\omega$  and quantum number  $l$  but with negative angular momentum  $-m$ . Thus the particle emission tends to carry away the angular momentum. For Boson fields, the factor in curly brackets in (3.4) is negative for  $\omega < m\Omega$ . However the fraction  $\Gamma_{jnlm}$  of the wave-packet that would have been absorbed by the black hole is also negative in this case because  $\omega < m\Omega$  is the condition for superradiance. In the limit that the temperature  $\kappa/2\pi$  is very low, the only particle emission occurs is an amount  $\mp \Gamma_{jnlm}$  in the modes for which  $\omega < m\Omega$ . This amount of particle creation is equal to that calculated by Starobinski [16] and Unruh [29], who considered only the final stationary Kerr solution and ignored the gravitational collapse.

One can treat a charged non-rotating black hole in a rather similar way. The behaviour of fields like the electromagnetic or gravitational fields which do not carry an electric charge will be the same as before except that the charge on the black will reduce the surface gravity  $k$  and hence the temperature of the black hole. Consider now the simple case of a massless charged scalar field  $\phi$  which obeys the minimally coupled wave equation

$$g^{ab}(V_a - ieA_a)(V_b - ieA_b)\phi = 0. \quad (3.5)$$

The phase of a solution  $p_\omega$  of the wave equation (3.5) is not gauge-invariant but the propagation vector  $ik_a = \nabla_a(\log p_\omega) - ieA_a$  is. In the geometric optics or WKB limit the vector  $k_a$  is null and propagates according to

$$k_{a;b}k^b = -eF_{ab}k^b. \quad (3.6)$$

An infinitesimal vector  $z^a$  will connect points with a “guage invariant” phase difference of  $ik_a z^a$ . If  $z^a$  is propagated along the integral curves of  $k^a$  according to

$$z_{;b}^a k^b = -eF_b^a z^b \quad (3.7)$$

$z^a$  will connect surfaces of constant guage invariant phase difference.

In the final stationary region one can choose a guage such that the electromagnetic potential  $A_a$  is stationary and vanishes on  $\mathcal{I}^+$ . In this guage the field equation (3.5) is separable and has solutions  $p_\omega$  with retarded time dependence  $e^{i\omega t}$ . Let  $x$  be a point on the event horizon in the final stationary region and let  $l^a$  and  $n^a$  be a pair of null vectors at  $x$ . As before, the vector  $-en^a$  will connect the event horizon with the surface of actual phase  $-\omega/\kappa (\log \varepsilon - \log C)$  of the solution  $p_\omega$ . However the guage invariant phase will be  $-\kappa^{-1}(\omega - e\Phi)(\log \varepsilon - \log C)$  where  $\Phi = K^a A_a$  is the electrostatic potential on the horizon and  $K^a$  is the time-translation Killing vector. Now propagate  $l^a$  like  $k^a$  in Eq. (3.6) back until it intersects a generator  $\lambda$  of  $\mathcal{I}^-$  at a point  $y$  and propagate  $n^a$  like  $z^a$  in Eq. (3.7) along the integral curve of  $l^a$ . With this propagation law, the vector  $-en^a$  will connect surfaces of constant guage invariant phase. Near  $\mathcal{I}^-$  one can use a different electromagnetic guage such that  $A^a$  is zero on  $\mathcal{I}^-$ . In this guage the phase of  $p_\omega^{(2)}$  along each generator of  $\mathcal{I}^-$  will have the form

$$-(\omega - e\phi)\kappa^{-1}\{\log(v_0 - v) - \log H\} \quad (3.8)$$

where  $H$  is a constant along each generator. This phase dependence gives the same thermal emission as before but with  $\omega$  replaced by  $\omega - e\Phi$ . Similar remarks apply about charge loss and superradiance. In the case that the black hole is both rotating and charged one can simply combine the above results.

#### 4. The Back-Reaction on the Metric

I now come to the difficult problem of the back-reaction of the particle creation on the metric and the consequent slow decrease of the mass of the black hole. At first sight it might seem that since all the time dependence of the metric in Fig. 4 is in the collapsing phase, all the particle creation must take place in the collapsing body just before the formation of the event horizon, and that an infinite number of created particles would hover just outside the event horizon, escaping to  $\mathcal{I}^+$  at a steady rate. This does not seem reasonable because it would involve the collapsing body knowing just when it was about to fall through the event horizon whereas the position of the event horizon is determined by the whole future history of the black hole and may be somewhat outside the apparent horizon, which is the only thing that can be determined locally [7].

Consider an observer falling through the horizon at some time after the collapse. He can set up a local inertial coordinate patch of radius  $\sim M$  centred

on the point where he crosses the horizon. He can pick a complete family  $\{h_\omega\}$  of solutions of the wave equations which obey the condition:

$$\frac{1}{2}i \int_S (h_{\omega_1} \bar{h}_{\omega_2;a} - \bar{h}_{\omega_2} h_{\omega_1;a}) d\Sigma^a = \delta(\omega_1 - \omega_2) \quad (4.1)$$

(where  $S$  is a Cauchy surface) and which have the approximate coordinate dependence  $e^{i\omega t}$  in the coordinate patch. This last condition determines the splitting into positive and negative frequencies and hence the annihilation and creation operators fairly well for modes  $h_\omega$  with  $\omega > M$  but not for those with  $\omega < M$ . Because the  $\{h_\omega\}$ , unlike the  $\{p_\omega\}$ , are continuous across the event horizon, they will also be continuous on  $\mathcal{I}^-$ . It is the discontinuity in the  $\{p_\omega\}$  on  $\mathcal{I}^-$  at  $v=v_0$  which is responsible for creating an infinite total number of particles in each mode.  $p_\omega$  by producing an  $(\omega')^{-1}$  tail in the Fourier transforms of the  $\{p_\omega\}$  at large negative frequencies  $\omega'$ . On the other hand, the  $\{h_\omega\}$  for  $\omega > M$  will have very small negative frequency components on  $\mathcal{I}^-$ . This means that the observer at the event horizon will see few particles with  $\omega > M$ . He will not be able to detect particles with  $\omega < M$  because they will have a wavelength bigger than his particle detector which must be smaller than  $M$ . As described in the introduction, there will be an indeterminacy in the energy density of order  $M^{-4}$  corresponding to the indeterminacy in the particle number for these modes.

The above discussion shows that the particle creation is really a global process and is not localised in the collapse: an observer falling through the event horizon would not see an infinite number of particles coming out from the collapsing body. Because it is a non-local process, it is probably not reasonable to expect to be able to form a local energy-momentum tensor to describe the back-reaction of the particle creation on the metric. Rather, the negative energy density needed to account for the decrease in the area of the horizon, should be thought of as arising from the indeterminacy of order of  $M^{-4}$  of the local energy density at the horizon. Equivalently, one can think of the area decrease as resulting from the fact that quantum fluctuations of the metric will cause the position and the very concept of the event horizon to be somewhat indeterminate.

Although it is probably not meaningful to talk about the local energy-momentum of the created particles, one may still be able to define the total energy flux over a suitably large surface. The problem is rather analogous to that of defining gravitational energy in classical general relativity: there are a number of different energy-momentum pseudo-tensors, none of which have any invariant local significance, but which all agree when integrated over a sufficiently large surface. In the particle case there are similarly a number of different expressions one can use for the renormalised energy-momentum tensor. The energy-momentum tensor for a classical field  $\phi$  is

$$T_{ab} = \phi_{;a}\phi_{;b} - \frac{1}{2}g_{ab}\phi^{cd}\phi_{;c}\phi_{;d}. \quad (4.2)$$

If one takes this expression over into the quantum theory and regards the  $\phi$ 's as operators one obtains a divergent result because there is a creation operator for each mode to the right of an annihilation operator. One therefore has to subtract out the divergence in some way. Various methods have been proposed for this (e.g. [30]) but they all seem a bit ad hoc. However, on the analogy of the pseudo-tensor, one would hope that the different renormalisations would all give the

same integrated fluxes. This is indeed the case in the final quasi-stationary region: all renormalised energy-momentum operators  $T_{ab}$  which obey the conservation equations  $T_{;b}^{ab}=0$ , which are stationary i.e. which have zero Lie derivative with respect to the time translation Killing vector  $K^a$  and which agree near  $\mathcal{I}^+$  will give the same fluxes of energy and angular momentum over any surface of constant  $r$  outside the event horizon. It is therefore sufficient to evaluate the energy flux near  $\mathcal{I}^+$ : by the conservation equations this will be equal to the energy flux out from the event horizon. Near  $\mathcal{I}^+$  the obvious way to renormalise the energy-momentum operator is to normal order the expression (4.2) with respect to positive and negative frequencies defined by the time-translation Killing vector  $K^a$  of the final quasi-stationary state. Near the event horizon normal ordering with respect to  $K^a$  cannot be the correct way to renormalise the energy-momentum operator since the normal-ordered operator diverges at the horizon. However it still gives the same energy outflow across any surface of constant  $r$ . A renormalised operator which was regular at the horizon would have to violate the weak energy condition by having negative energy density. This negative energy density is not observable locally.

In order to evaluate the normal ordered operator one wants to choose the  $\{q_i\}$  which describe waves crossing the event horizon, to be positive frequency with respect to the time parameter defined by  $K^a$  along the generators of the horizon in the final quasi-stationary state. The condition on the  $\{q_i\}$  in the time-dependent collapse phase is not determined but this should not affect wave packets on the horizon at late times. If one makes up wave-packets  $\{q_{jn}\}$  like the  $\{p_{jn}\}$ , one finds that a fraction  $\Gamma_{jn}$  penetrates through the potential barrier around the black hole and gets out to  $\mathcal{I}^-$  with the same frequency  $\omega$  that it had on the horizon. This produces a  $\delta(\omega - \omega')$  behaviour in  $\gamma_{jn\omega'}$ . The remaining fraction  $1 - \Gamma_{jn}$  of the wave-packet is reflected back by the potential barrier and passes through the collapsing body and out onto  $\mathcal{I}^-$ . Here it will have a similar form to  $p_{jn}^{(2)}$ . Thus for large  $\omega'$ ,

$$|\gamma_{jn\omega'}^{(2)}| = \exp(\pi\omega\kappa^{-1}) |\eta_{jn\omega'}^{(2)}|. \quad (4.3)$$

By a similar argument to that used in Section (2) one would conclude that the number of particles crossing the event horizon in a wave-packet mode peaked at late times would be

$$(1 - \Gamma_{jn}) \{\exp(2\pi\omega\kappa^{-1}) - 1\}^{-1}. \quad (4.4)$$

For a given frequency  $\omega$ , i.e. a given value of  $j$ , the absorption fraction  $\Gamma_{jn}$  goes to zero as the angular quantum number  $l$  increases because of the centrifugal barrier. Thus at first sight it might seem that each wave-packet mode of high  $l$  value would contain

$$\{\exp(2\pi\omega\kappa^{-1}) - 1\}^{-1}$$

particles and that the total rate of particles and energy crossing the event horizon would be infinite. This calculation would, of course, be inconsistent with the result obtained above that an observer crossing the event horizon would see only a finite small energy density of order  $M^{-4}$ . The reason for this discrepancy seems to be that the wave-packets  $\{p_{jn}\}$  and  $\{q_{jn}\}$  provide a complete basis for solutions

of the wave equation only in the region outside the event horizon and not actually on the event horizon itself. In order to calculate the particle flux over the horizon one therefore has to calculate the flux over some surface just outside the horizon and take the limit as the surface approaches the horizon.

To perform this calculation it is convenient to define new wave-packets  $x_{jn} = p_{jn}^{(2)} + q_{jn}^{(2)}$  which represent the part of  $p_{jn}$  and  $q_{jn}$  which passes through the collapsing body and  $y_{jn} = p_{jn}^{(1)} + q_{jn}^{(1)}$  which represents the part of  $p_{jn}$  and  $q_{jn}$  which propagates out to  $\mathcal{I}^-$  through the quasi-stationary metric of the final black hole. In the initial vacuum state the  $\{y_{jn}\}$  modes will not contain any particles but each  $x_{jn}$  mode will contain  $\{\exp(2\pi\omega\kappa^{-1}) - 1\}^{-1}$  particles. These particles will appear to leave the collapsing body just outside the event horizon and will propagate radially outwards. A fraction  $\Gamma_{jn}$  will penetrate through the potential barrier peaked at  $r=3M$  and will escape to  $\mathcal{I}^+$  where they will constitute the thermal emission of the black hole. The remaining fraction  $1-\Gamma_{jn}$  will be reflected back by the potential barrier and will cross the event horizon. Thus the net particle flux across a surface of constant  $r$  just outside the horizon will be  $\Gamma_{jn}$  directed outwards.

I shall now show that using the normal ordered energy momentum operator, the average energy flux across a surface of constant  $r$  between retarded times  $u_1$  and  $u_2$

$$(u_2 - u_1)^{-1} \int_{u_1}^{u_2} \langle 0_- | T_{ab} | 0_- \rangle K^a d\Sigma^b \quad (4.5)$$

is directed outwards and is equal to the energy flux for the thermal emission from a hot body. Because the  $\{y_{jn}\}$  contain no negative frequencies on  $\mathcal{I}^-$ , they will not make any contribution to the expectation value (4.5) of the normal ordered energy-momentum operator. Let

$$x_{jn} = \int_0^\infty (\zeta_{j\omega'} f_{\omega'} + \xi_{j\omega'} \bar{f}_{\omega'}) d\omega'. \quad (4.6)$$

Near  $\mathcal{I}^+$

$$x_{jn} = (\Gamma_{jn})^{\frac{1}{2}} p_{jn}. \quad (4.7)$$

Thus

$$(4.5) = (u_2 - u_1)^{-1} \operatorname{Re} \left\{ \sum_{j,n} \sum_{j'',n''} \int_0^\infty \int_{u_1}^{u_2} \omega \omega'' \Gamma_{jn}^{\frac{1}{2}} p_{jn} \bar{\xi}_{j\omega'} \right. \\ \cdot \left. (\bar{\Gamma}_{j''n''}^{\frac{1}{2}} \bar{p}_{j''n''} \xi_{j''n''\omega'} - \Gamma_{j''n''}^{\frac{1}{2}} p_{j''n''} \zeta_{j''n''\omega'}) d\omega' du \right\} \quad (4.8)$$

where  $\omega$  and  $\omega''$  are the frequencies of the wave-packets  $p_{jn}$  and  $p_{j''n''}$  respectively. In the limit  $u_2 - u_1$  tends to infinity, the second term in the integrand in (4.8) will integrate out and the first term will contribute only for  $(j'', n'') = (j, n)$ . By arguments similar to those used in Section 2,

$$\int_0^\infty |\xi_{j\omega'}|^2 d\omega' = \{\exp(2\pi\omega\kappa^{-1}) - 1\}^{-1}. \quad (4.9)$$

Therefore

$$(4.5) = \int_0^\infty \Gamma_\omega \omega \{\exp(2\pi\omega\kappa^{-1}) - 1\}^{-1} d\omega \quad (4.10)$$

where  $\Gamma_\omega = \lim_{n \rightarrow \infty} \Gamma_{jn}$  is the fraction of wave-packet of frequency that would be absorbed by the black hole. The energy flux (4.10) corresponds exactly to the rate of thermal emission calculated in Section 2. Any renormalised energy momentum

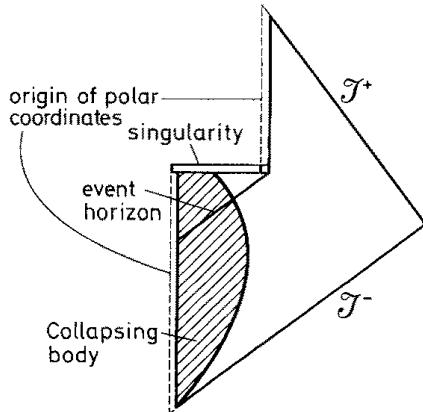


Fig. 5. The Penrose diagram for a gravitational collapse followed by the slow evaporation and eventual disappearance of the black hole, leaving empty space with no singularity at the origin

operator which agrees with the normal ordered operator near  $\mathcal{I}^+$ , which obeys the conservation equations, and which is stationary in the final quasi-stationary region will give the same energy flux over any surface of constant  $r$ . Thus it will give positive energy flux out across the event horizon or, equivalently, a negative energy flux in across the event horizon.

This negative energy flux will cause the area of the event horizon to decrease and so the black hole will not, in fact, be in a stationary state. However, as long as the mass of the black hole is large compared to the Planck mass  $10^{-5}$  g, the rate of evolution of the black hole will be very slow compared to the characteristic time for light to cross the Schwarzschild radius. Thus it is a reasonable approximation to describe the black hole by a sequence of stationary solutions and to calculate the rate of particle emission in each solution. Eventually, when the mass of the black hole is reduced to  $10^{-5}$  g, the quasi-stationary approximation will break down. At this point, one cannot continue to use the concept of a classical metric. However, the total mass or energy remaining in the system is very small. Thus, provided the black hole does not evolve into a negative mass naked singularity there is not much it can do except disappear altogether. The baryons or leptons that formed the original collapsing body cannot reappear because all their rest mass energy has been carried away by the thermal radiation. It is tempting to speculate that this might be the reason why the universe now contains so few baryons compared to photons: the universe might have started out with baryons only, and no radiation. Most of the baryons might have fallen into small black holes which then evaporated giving back the rest mass energy of baryons in the form of radiation, but not the baryons themselves.

The Penrose diagram of a black hole which evaporates and leaves only empty space is shown in Fig. 5. The horizontal line marked "singularity" is really a region where the radius of curvature is of the order the Planck length. The matter that runs into this region might reemerge in another universe or it might even reemerge in our universe through the upper vertical line thus creating a naked singularity of negative mass.

### References

1. Isham, C. J.: Preprint (1973)
2. Ashtekar, A., Geroch, R. P.: Quantum theory of gravity (preprint 1973)
3. Penrose, R.: Phys. Rev. Lett. **14**, 57—59 (1965)
4. Hawking, S. W.: Proc. Roy. Soc. Lond. A **300**, 187—20 (1967)
5. Hawking, S. W., Penrose, R.: Proc. Roy. Soc. Lond. A **314**, 529—548 (1970)
6. Hawking, S. W., Ellis, G. F. R.: The large scale structure of space-time. London: Cambridge University Press 1973
7. Hawking, S. W.: The event horizon. In: Black holes. Ed. C. M. DeWitt, B. S. DeWitt. New York: Gordon and Breach 1973
8. Bardeen, J. M., Carter, B., Hawking, S. W.: Commun. math. Phys. **31**, 161—170 (1973)
9. Hawking, S. W.: Mon. Not. Roy. astr. Soc. **152**, 75—78 (1971)
10. Carr, B. J., Hawking, S. W.: Monthly Notices Roy. Astron. Soc. **168**, 399—415 (1974)
11. Hagedorn, R.: Astron. Astrophys. **5**, 184 (1970)
12. Hawking, S. W.: Commun. math. Phys. **25**, 152—166 (1972)
13. Carter, B.: Black hole equilibrium states. In: Black holes. Ed. C. M. DeWitt, B. S. DeWitt. New York: Gordon and Breach 1973
14. Misner, C. W.: Bull. Amer. Phys. Soc. **17**, 472 (1972)
15. Press, W. M., Teukolsky, S. A.: Nature **238**, 211 (1972)
16. Starobinsky, A. A.: Zh. E.T.F. **64**, 48 (1973)
17. Starobinsky, A. A., Churilov, S. M.: Zh. E.T.F. **65**, 3 (1973)
18. Bjorken, T. D., Drell, S. D.: Relativistic quantum mechanics. New York: McGraw Hill 1965
19. Beckenstein, J. D.: Phys. Rev. D. **7**, 2333—2346 (1973)
20. Beckenstein, J. D.: Phys. Rev. D. **9**,
21. Penrose, R.: Phys. Rev. Lett. **10**, 66—68 (1963)
22. Sachs, R. K.: Proc. Roy. Soc. Lond. A **270**, 103 (1962)
23. Eardley, D., Sachs, R. K.: J. Math. Phys. **14** (1973)
24. Schmidt, B. G.: Commun. Math. Phys. **36**, 73—90 (1974)
25. Bondi, H., van der Burg, M. G. J., Metzner, A. W. K.: Proc. Roy. Soc. Lond. A **269**, 21 (1962)
26. Carter, B.: Commun. math. Phys. **10**, 280—310 (1968)
27. Teukolsky, S. A.: Ap. J. **185**, 635—647 (1973)
28. Unruh, W.: Phys. Rev. Lett. **31**, 1265 (1973)
29. Unruh, W.: Phys. Rev. D. **10**, 3194—3205 (1974)
30. Zeldovich, Ya. B., Starobinsky, A. A.: Zh. E.T.F. **61**, 2161 (1971), JETP **34**, 1159 (1972)

Communicated by J. Ehlers

S. W. Hawking  
 California Institute of Technology  
 W. K. Kellogg Radiation Lab. 106-38  
 Pasadena, California 91125, USA

## Inflationary universe: A possible solution to the horizon and flatness problems

Alan H. Guth\*

*Stanford Linear Accelerator Center, Stanford University, Stanford, California 94305*

(Received 11 August 1980)

The standard model of hot big-bang cosmology requires initial conditions which are problematic in two ways: (1) The early universe is assumed to be highly homogeneous, in spite of the fact that separated regions were causally disconnected (horizon problem); and (2) the initial value of the Hubble constant must be fine tuned to extraordinary accuracy to produce a universe as flat (i.e., near critical mass density) as the one we see today (flatness problem). These problems would disappear if, in its early history, the universe supercooled to temperatures 28 or more orders of magnitude below the critical temperature for some phase transition. A huge expansion factor would then result from a period of exponential growth, and the entropy of the universe would be multiplied by a huge factor when the latent heat is released. Such a scenario is completely natural in the context of grand unified models of elementary-particle interactions. In such models, the supercooling is also relevant to the problem of monopole suppression. Unfortunately, the scenario seems to lead to some unacceptable consequences, so modifications must be sought.

### I. INTRODUCTION: THE HORIZON AND FLATNESS PROBLEMS

The standard model of hot big-bang cosmology relies on the assumption of initial conditions which are very puzzling in two ways which I will explain below. The purpose of this paper is to suggest a modified scenario which avoids both of these puzzles.

By "standard model," I refer to an adiabatically expanding radiation-dominated universe described by a Robertson-Walker metric. Details will be given in Sec. II.

Before explaining the puzzles, I would first like to clarify my notion of "initial conditions." The standard model has a singularity which is conventionally taken to be at time  $t=0$ . As  $t \rightarrow 0$ , the temperature  $T \rightarrow \infty$ . Thus, no initial-value problem can be defined at  $t=0$ . However, when  $T$  is of the order of the Planck mass ( $M_P \equiv 1/\sqrt{G} = 1.22 \times 10^{19}$  GeV)<sup>1</sup> or greater, the equations of the standard model are undoubtedly meaningless, since quantum gravitational effects are expected to become essential. Thus, within the scope of our knowledge, it is sensible to begin the hot big-bang scenario at some temperature  $T_0$  which is comfortably below  $M_P$ ; let us say  $T_0 = 10^{17}$  GeV. At this time one can take the description of the universe as a set of initial conditions, and the equations of motion then describe the subsequent evolution. Of course, the equation of state for matter at these temperatures is not really known, but one can make various hypotheses and pursue the consequences.

In the standard model, the initial universe is taken to be homogeneous and isotropic, and filled with a gas of effectively massless particles in thermal equilibrium at temperature  $T_0$ . The initial value of the Hubble expansion "constant"  $H$  is taken to be  $H_0$ , and the model universe is then

completely described.

Now I can explain the puzzles. The first is the well-known horizon problem.<sup>2-4</sup> The initial universe is assumed to be homogeneous, yet it consists of at least  $\sim 10^{83}$  separate regions which are causally disconnected (i.e., these regions have not yet had time to communicate with each other via light signals).<sup>5</sup> (The precise assumptions which lead to these numbers will be spelled out in Sec. II.) Thus, one must assume that the forces which created these initial conditions were capable of violating causality.

The second puzzle is the flatness problem. This puzzle seems to be much less celebrated than the first, but it has been stressed by Dicke and Peebles.<sup>6</sup> I feel that it is of comparable importance to the first. It is known that the energy density  $\rho$  of the universe today is near the critical value  $\rho_{cr}$  (corresponding to the borderline between an open and closed universe). One can safely assume that<sup>7</sup>

$$0.01 < \Omega_p < 10, \quad (1.1)$$

where

$$\Omega \equiv \rho / \rho_{cr} = (8\pi/3)G\rho/H^2, \quad (1.2)$$

and the subscript  $p$  denotes the value at the present time. Although these bounds do not appear at first sight to be remarkably stringent, they, in fact, have powerful implications. The key point is that the condition  $\Omega \approx 1$  is unstable. Furthermore, the only time scale which appears in the equations for a radiation-dominated universe is the Planck time,  $1/M_P = 5.4 \times 10^{-44}$  sec. A typical closed universe will reach its maximum size on the order of this time scale, while a typical open universe will dwindle to a value of  $\rho$  much less than  $\rho_{cr}$ . A universe can survive  $\sim 10^{10}$  years only by extreme fine tuning of the initial values of  $\rho$  and  $H$ , so that  $\rho$  is very near  $\rho_{cr}$ . For the initial conditions taken at

$T_0 = 10^{17}$  GeV, the value of  $H_0$  must be fine tuned to an accuracy of one part in  $10^{55}$ . In the standard model this incredibly precise initial relationship must be assumed without explanation. (For any reader who is not convinced that there is a real problem here, variations of this argument are given in the Appendix.)

The reader should not assume that these incredible numbers are due merely to the rather large value I have taken for  $T_0$ . If I had chosen a modest value such as  $T_0 = 1$  MeV, I would still have concluded that the “initial” universe consisted of at least  $\sim 10^{22}$  causally disconnected regions, and that the initial value of  $H_0$  was fine tuned to one part in  $10^{15}$ . These numbers are much smaller than the previous set, but they are still very impressive.

Of course, any problem involving the initial conditions can always be put off until we understand the physics of  $T \gtrsim M_P$ . However, it is the purpose of this paper to show that these puzzles might be obviated by a scenario for the behavior of the universe at temperatures well below  $M_P$ .

The paper is organized as follows. The assumptions and basic equations of the standard model are summarized in Sec. II. In Sec. III, I describe the inflationary universe scenario, showing how it can eliminate the horizon and flatness problems. The scenario is discussed in the context of grand models in Sec. IV, and comments are made concerning magnetic monopole suppression. In Sec. V I discuss briefly the key undesirable feature of the scenario: the inhomogeneities produced by the random nucleation of bubbles. Some vague ideas which might alleviate these difficulties are mentioned in Sec. VI.

## II. THE STANDARD MODEL OF THE VERY EARLY UNIVERSE

In this section I will summarize the basic equations of the standard model, and I will spell out the assumptions which lead to the statements made in the Introduction.

The universe is assumed to be homogeneous and isotropic, and is therefore described by the Robertson-Walker metric<sup>8</sup>:

$$d\tau^2 = dt^2 - R^2(t) \left[ \frac{dr^2}{1 - kr^2} + r^2(d\theta^2 + \sin^2\theta d\phi^2) \right], \quad (2.1)$$

where  $k = +1, -1$ , or  $0$  for a closed, open, or flat universe, respectively. It should be emphasized that any value of  $k$  is possible, but by convention  $r$  and  $R(t)$  are rescaled so that  $k$  takes on one of the three discrete values. The evolution of  $R(t)$  is governed by the Einstein equations

$$\ddot{R} = -\frac{4\pi}{3}G(\rho + 3p)R, \quad (2.2a)$$

$$H^2 + \frac{k}{R^2} = \frac{8\pi}{3}G\rho, \quad (2.2b)$$

where  $H \equiv \dot{R}/R$  is the Hubble “constant” (the dot denotes the derivative with respect to  $t$ ). Conservation of energy is expressed by

$$\frac{d}{dt}(\rho R^3) = -p \frac{d}{dt}(R^3), \quad (2.3)$$

where  $p$  denotes the pressure. In the standard model one also assumes that the expansion is adiabatic, in which case

$$\frac{d}{dt}(sR^3) = 0, \quad (2.4)$$

where  $s$  is the entropy density.

To determine the evolution of the universe, the above equations must be supplemented by an equation of state for matter. It is now standard to describe matter by means of a field theory, and at high temperatures this means that the equation of state is to a good approximation that of an ideal quantum gas of massless particles. Let  $N_b(T)$  denote the number of bosonic spin degrees of freedom which are effectively massless at temperature  $T$  (e.g., the photon contributes two units to  $N_b$ ); and let  $N_f(T)$  denote the corresponding number for fermions (e.g., electrons and positrons together contribute four units). Provided that  $T$  is not near any mass thresholds, the thermodynamic functions are given by

$$\rho = 3p = \frac{\pi^2}{30}\mathfrak{N}(T)T^4, \quad (2.5)$$

$$s = \frac{2\pi^2}{45}\mathfrak{N}(T)T^3, \quad (2.6)$$

$$n = \frac{\zeta(3)}{\pi^2}\mathfrak{N}'(T)T^3, \quad (2.7)$$

where

$$\mathfrak{N}(T) = N_b(T) + \frac{7}{8}N_f(T), \quad (2.8)$$

$$\mathfrak{N}'(T) = N_b(T) + \frac{3}{4}N_f(T). \quad (2.9)$$

Here  $n$  denotes the particle number density, and  $\zeta(3) = 1.20206\dots$  is the Riemann zeta function.

The evolution of the universe is then found by rewriting (2.2b) solely in terms of the temperature. Again assuming that  $T$  is not near any mass thresholds, one finds

$$\left(\frac{\dot{T}}{T}\right)^2 + \epsilon(T)T^2 = \frac{4\pi^3}{45}G\mathfrak{N}(T)T^4, \quad (2.10)$$

where

$$\epsilon(T) = \frac{k}{R^2 T^2} = k \left[ \frac{2\pi^2}{45} \frac{\mathfrak{N}(T)}{S} \right]^{2/3}, \quad (2.11)$$

where  $S \equiv R^3 s$  denotes the total entropy in a volume

specified by the radius of curvature  $R$ .

Since  $S$  is conserved, its value in the early universe can be determined (or at least bounded) by current observations. Taking  $\rho < 10\rho_{\text{cr}}$  today, it follows that today

$$\left| \frac{k}{R^2} \right| < 9H^2. \quad (2.12)$$

From now on I will take  $k = \pm 1$ ; the special case  $k = 0$  is still included as the limit  $R \rightarrow \infty$ . Then today  $R > \frac{1}{3}H^{-1} \sim 3 \times 10^9$  years. Taking the present photon temperature  $T_\gamma$  as 2.7°K, one then finds that the photon contribution to  $S$  is bounded by

$$S_\gamma > 3 \times 10^{85}. \quad (2.13)$$

Assuming that there are three species of massless neutrinos ( $e$ ,  $\mu$ , and  $\tau$ ), all of which decouple at a time when the other effectively massless particles are the electrons and photons, then  $S_\nu = 21/22S_\gamma$ . Thus,

$$S > 10^{86} \quad (2.14)$$

and

$$|\epsilon| < 10^{-58}\pi^{2/3}. \quad (2.15)$$

But then

$$\left| \frac{\rho - \rho_{\text{cr}}}{\rho} \right| = \frac{45}{4\pi^3} \frac{M_P^2}{\pi T^2} |\epsilon| < 3 \times 10^{-59}\pi^{-1/3}(M_P/T)^2. \quad (2.16)$$

Taking  $T = 10^{17}$  GeV and  $\pi \sim 10^2$  (typical of grand unified models), one finds  $|\rho - \rho_{\text{cr}}|/\rho < 10^{-55}$ . This is the flatness problem.

The  $\epsilon T^2$  term can now be deleted from (2.10), which is then solved (for temperatures higher than all particle masses) to give

$$T^2 = \frac{M_P}{2\gamma t}, \quad (2.17)$$

where  $\gamma^2 = (4\pi^3/45)\pi$ . (For the minimal  $SU_5$  grand unified model,  $N_b = 82$ ,  $N_f = 90$ , and  $\gamma = 21.05$ .)

Conservation of entropy implies  $RT = \text{constant}$ , so  $R \propto t^{1/2}$ . A light pulse beginning at  $t = 0$  will have traveled by time  $t$  a physical distance

$$l(t) = R(t) \int_0^t dt' R^{-1}(t') = 2t, \quad (2.18)$$

and this gives the physical horizon distance. This horizon distance is to be compared with the radius  $L(t)$  of the region at time  $t$  which will evolve into our currently observed region of the universe. Again using conservation of entropy,

$$L(t) = [s_p/s(t)]^{1/3} L_p, \quad (2.19)$$

where  $s_p$  is the present entropy density and  $L_p \sim 10^{10}$  years is the radius of the currently observed region of the universe. One is interested in the

ratio of volumes, so

$$\begin{aligned} \frac{l^3}{L^3} &= \frac{11}{43} \left( \frac{45}{4\pi^3} \right)^{3/2} \pi^{-1/2} \left( \frac{M_P}{L_p T_\gamma T} \right)^3 \\ &= 4 \times 10^{-89} \pi^{-1/2} (M_P/T)^3. \end{aligned} \quad (2.20)$$

Taking  $\pi \sim 10^2$  and  $T_0 = 10^{17}$  GeV, one finds  $l_0^3/L_0^3 = 10^{-83}$ . This is the horizon problem.

### III. THE INFLATIONARY UNIVERSE

In this section I will describe a scenario which is capable of avoiding the horizon and flatness problems.

From Sec. II one can see that both problems could disappear if the assumption of adiabaticity were grossly incorrect. Suppose instead that

$$S_p = Z^3 S_0, \quad (3.1)$$

where  $S_p$  and  $S_0$  denote the present and initial values of  $R^3 s$ , and  $Z$  is some large factor.

Let us look first at the flatness problem. Given (3.1), the right-hand side (RHS) of (2.16) is multiplied by a factor of  $Z^2$ . The "initial" value (at  $T_0 = 10^{17}$  GeV) of  $|\rho - \rho_{\text{cr}}|/\rho$  could be of order unity, and the flatness problem would be obviated, if

$$Z > 3 \times 10^{27}. \quad (3.2)$$

Now consider the horizon problem. The RHS of (2.19) is multiplied by  $Z^{-1}$ , which means that the length scale of the early universe, at any given temperature, was smaller by a factor of  $Z$  than had been previously thought. If  $Z$  is sufficiently large, then the initial region which evolved into our observed region of the universe would have been smaller than the horizon distance at that time. To see how large  $Z$  must be, note that the RHS of (2.20) is multiplied by  $Z^3$ . Thus, if

$$Z > 5 \times 10^{27}, \quad (3.3)$$

then the horizon problem disappears. (It should be noted that the horizon will still exist; it will simply be moved out to distances which have not been observed.)

It is not surprising that the RHS's of (3.2) and (3.3) are approximately equal, since they both correspond roughly to  $S_0$  of order unity.

I will now describe a scenario, which I call the inflationary universe, which is capable of such a large entropy production.

Suppose the equation of state for matter (with all chemical potentials set equal to zero) exhibits a first-order phase transition at some critical temperature  $T_c$ . Then as the universe cools through the temperature  $T_c$ , one would expect bubbles of the low-temperature phase to nucleate and grow. However, suppose the nucleation rate for this phase transition is rather low. The universe will

continue to cool as it expands, and it will then supercool in the high-temperature phase. Suppose that this supercooling continues down to some temperature  $T_s$ , many orders of magnitude below  $T_c$ . When the phase transition finally takes place at temperature  $T_s$ , the latent heat is released. However, this latent heat is characteristic of the energy scale  $T_c$ , which is huge relative to  $T_s$ . The universe is then reheated to some temperature  $T_r$  which is comparable to  $T_c$ . The entropy density is then increased by a factor of roughly  $(T_r/T_s)^3$  (assuming that the number  $\pi$  of degrees of freedom for the two phases are comparable), while the value of  $R$  remains unchanged. Thus,

$$Z \approx T_r/T_s. \quad (3.4)$$

If the universe supercools by 28 or more orders of magnitude below some critical temperature, the horizon and flatness problems disappear.

In order for this scenario to work, it is necessary for the universe to be essentially devoid of any strictly conserved quantities. Let  $n$  denote the density of some strictly conserved quantity, and let  $r = n/s$  denote the ratio of this conserved quantity to entropy. Then  $r_b = Z^3 r_0 < 10^{-24} r_0$ . Thus, only an absurdly large value for the initial ratio would lead to a measurable value for the present ratio. Thus, if baryon number were exactly conserved, the inflationary model would be untenable. However, in the context of grand unified models, baryon number is not exactly conserved. The net baryon number of the universe is believed to be created by  $CP$ -violating interactions at a temperature of  $10^{13}$ – $10^{14}$  GeV.<sup>9</sup> Thus, provided that  $T_c$  lies in this range or higher, there is no problem. The baryon production would take place after the reheating. (However, strong constraints are imposed on the entropy which can be generated in any phase transition with  $T_c \ll 10^{14}$  GeV, in particular, the Weinberg-Salam phase transition.<sup>36</sup>)

Let us examine the properties of the supercooling universe in more detail. Note that the energy density  $\rho(T)$ , given in the standard model by (2.5), must now be modified. As  $T \rightarrow 0$ , the system is cooling not toward the true vacuum, but rather toward some metastable false vacuum with an energy density  $\rho_0$  which is necessarily higher than that of the true vacuum. Thus, to a good approximation (ignoring mass thresholds)

$$\rho(T) = \frac{\pi^2}{30} \mathfrak{U}(T) T^4 + \rho_0. \quad (3.5)$$

Perhaps a few words should be said concerning the zero point of energy. Classical general relativity couples to an energy-momentum tensor of matter,  $T_{\mu\nu}$ , which is necessarily (covariantly) conserved. When matter is described by a field

theory, the form of  $T_{\mu\nu}$  is determined by the conservation requirement up to the possible modification

$$T_{\mu\nu} \rightarrow T_{\mu\nu} + \lambda g_{\mu\nu}, \quad (3.6)$$

for any constant  $\lambda$ . ( $\lambda$  cannot depend on the values of the fields, nor can it depend on the temperature or the phase.) The freedom to introduce the modification (3.6) is identical to the freedom to introduce a cosmological constant into Einstein's equations. One can always choose to write Einstein's equations without an explicit cosmological term; the cosmological constant  $\Lambda$  is then defined by

$$\langle 0 | T_{\mu\nu} | 0 \rangle = \Lambda g_{\mu\nu}, \quad (3.7)$$

where  $|0\rangle$  denotes the true vacuum.  $\Lambda$  is identified as the energy density of the vacuum, and, in principle, there is no reason for it to vanish. Empirically  $\Lambda$  is known to be very small ( $|\Lambda| < 10^{-46}$  GeV<sup>4</sup>)<sup>10</sup> so I will take its value to be zero.<sup>11</sup> The value of  $\rho_0$  is then necessarily positive and is determined by the particle theory.<sup>12</sup> It is typically of  $O(T_c^4)$ .

Using (3.5), Eq. (2.10) becomes

$$\left( \frac{\dot{T}}{T} \right)^2 = \frac{4\pi^3}{45} G \mathfrak{U}(T) T^4 - \epsilon(T) T^2 + \frac{8\pi}{3} G \rho_0. \quad (3.8)$$

This equation has two types of solutions, depending on the parameters. If  $\epsilon > \epsilon_0$ , where

$$\epsilon_0 = \frac{8\pi^2 \sqrt{30}}{45} G \sqrt{\mathfrak{U} \rho_0}, \quad (3.9)$$

then the expansion of the universe is halted at a temperature  $T_{\min}$  given by

$$T_{\min}^4 = \frac{30\rho_0}{\pi^2} \left\{ \frac{\epsilon}{\epsilon_0} + \left[ \left( \frac{\epsilon}{\epsilon_0} \right)^2 - 1 \right]^{1/2} \right\}^2, \quad (3.10)$$

and then the universe contracts again. Note that  $T_{\min}$  is of  $O(T_c)$ , so this is not the desired scenario. The case of interest is  $\epsilon < \epsilon_0$ , in which case the expansion of the universe is unchecked. [Note that  $\epsilon_0 \sim \sqrt{\mathfrak{U}} T_c^2 / M_p^2$  is presumably a very small number. Thus  $0 < \epsilon < \epsilon_0$  (a closed universe) seems unlikely, but  $\epsilon < 0$  (an open universe) is quite plausible.] Once the temperature is low enough for the  $\rho_0$  term to dominate over the other two terms on the RHS of (3.8), one has

$$T(t) \approx \text{const} \times e^{-\chi t}, \quad (3.11)$$

where

$$\chi^2 = \frac{8\pi}{3} G \rho_0. \quad (3.12)$$

Since  $RT = \text{const}$ , one has<sup>13</sup>

$$R(t) = \text{const} \times e^{\chi t}. \quad (3.13)$$

The universe is expanding exponentially, in a false

vacuum state of energy density  $\rho_0$ . The Hubble constant is given by  $H = \dot{R}/R = \chi$ . (More precisely,  $H$  approaches  $\chi$  monotonically from above. This behavior differs markedly from the standard model, in which  $H$  falls as  $t^{-1}$ .)

The false vacuum state is Lorentz invariant, so  $T_{\mu\nu} = \rho_0 g_{\mu\nu}$ . It follows that  $p = -\rho_0$ , the pressure is negative. This negative pressure allows for the conservation of energy, Eq. (2.3). From the second-order Einstein equation (2.2a), it can be seen that the negative pressure is also the driving force behind the exponential expansion.

The Lorentz invariance of the false vacuum has one other consequence: The metric described by (3.13) (with  $k=0$ ) does not single out a comoving frame. The metric is invariant under an  $O(4,1)$  group of transformations, in contrast to the usual Robertson-Walker invariance of  $O(4)$ .<sup>14</sup> It is known as the de Sitter metric, and it is discussed in the standard literature.<sup>15</sup>

Now consider the process of bubble formation in a Robertson-Walker universe. The bubbles form randomly, so there is a certain nucleation rate  $\lambda(t)$ , which is the probability per (physical) volume per time that a bubble will form in any region which is still in the high-temperature phase. I will idealize the situation slightly and assume that the bubbles start at a point and expand at the speed of light. Furthermore, I neglect  $k$  in the metric, so  $d^2 = dt^2 - R^2(t)d\vec{x}^2$ .

I want to calculate  $p(t)$ , the probability that any given point remains in the high-temperature phase at time  $t$ . Note that the distribution of bubbles is totally uncorrelated except for the exclusion principle that bubbles do not form inside of bubbles. This exclusion principle causes no problem because one can imagine fictitious bubbles which form inside the real bubbles with the same nucleation rate  $\lambda(t)$ . With all bubbles expanding at the speed of light, the fictitious bubbles will be forever inside the real bubbles and will have no effect on  $p(t)$ . The distribution of all bubbles, real and fictitious, is then totally uncorrelated.

$p(t)$  is the probability that there are no bubbles which engulf a given point in space. But the number of bubbles which engulf a given point is a Poisson-distributed variable, so  $p(t) = \exp[-\bar{N}(t)]$ , where  $\bar{N}(t)$  is the expectation value of the number of bubbles engulfing the point. Thus<sup>16</sup>

$$p(t) = \exp\left[-\int_0^t dt_1 \lambda(t_1) R^3(t_1) V(t, t_1)\right], \quad (3.14)$$

where

$$V(t, t_1) = \frac{4\pi}{3} \left[ \int_{t_1}^t \frac{dt_2}{R(t_2)} \right]^3 \quad (3.15)$$

is the coordinate volume at time  $t$  of a bubble which

formed at time  $t_1$ .

I will now assume that the nucleation rate is sufficiently slow so that no significant nucleation takes place until  $T \ll T_c$ , when exponential growth has set in. I will further assume that by this time  $\lambda(t)$  is given approximately by the zero-temperature nucleation rate  $\lambda_0$ . One then has

$$p(t) = \exp\left[-\frac{t}{\tau} + O(1)\right], \quad (3.16)$$

where

$$\tau = \frac{3\chi^3}{4\pi\lambda_0}, \quad (3.17)$$

and  $O(1)$  refers to terms which approach a constant as  $xt \rightarrow \infty$ . During one of these time constants, the universe will expand by a factor

$$Z_\tau = \exp(\chi\tau) = \exp\left(\frac{3\chi^4}{4\pi\lambda_0}\right). \quad (3.18)$$

If the phase transition is associated with the expectation value of a Higgs field, then  $\lambda_0$  can be calculated using the method of Coleman and Callan.<sup>17</sup> The key point is that nucleation is a tunneling process, so that  $\lambda_0$  is typically very small. The Coleman-Callan method gives an answer of the form

$$\lambda_0 = A \rho_0 \exp(-B), \quad (3.19)$$

where  $B$  is a barrier penetration term and  $A$  is a dimensionless coefficient of order unity. Since  $Z_\tau$  is then an exponential of an exponential, one can very easily<sup>18,19,36</sup> obtain values as large as  $\log_{10} Z \approx 28$ , or even  $\log_{10} Z \approx 10^{10}$ .

Thus, if the universe reaches a state of exponential growth, it is quite plausible for it to expand and supercool by a huge number of orders of magnitude before a significant fraction of the universe undergoes the phase transition.

So far I have assumed that the early universe can be described from the beginning by a Robertson-Walker metric. If this assumption were really necessary, then it would be senseless to talk about "solving" the horizon problem; perfect homogeneity was assumed at the outset. Thus, I must now argue that the assumption can probably be dropped.

Suppose instead that the initial metric, and the distribution of particles, was rather chaotic. One would then expect that statistical effects would tend to thermalize the particle distribution on a local scale.<sup>20</sup> It has also been shown (in idealized circumstances) that anisotropies in the metric are damped out on the time scale of  $\sim 10^3$  Planck times.<sup>21</sup> The damping of inhomogeneities in the metric has also been studied,<sup>22</sup> and it is reasonable to expect such damping to occur. Thus, assuming that at least some region of the universe started at

temperatures high compared to  $T_c$ , one would expect that, by the time the temperature in one of these regions falls to  $T_c$ , it will be *locally* homogeneous, isotropic, and in thermal equilibrium. By locally, I am talking about a length scale  $\xi$  which is of course less than the horizon distance. It will then be possible to describe this local region of the universe by a Robertson-Walker metric, which will be accurate at distance scales small compared to  $\xi$ . When the temperature of such a region falls below  $T_c$ , the inflationary scenario will take place. The end result will be a huge region of space which is homogeneous, isotropic, and of nearly critical mass density. If  $Z$  is sufficiently large, this region can be bigger than (or much bigger than) our observed region of the universe.

#### IV. GRAND UNIFIED MODELS AND MAGNETIC MONOPOLE PRODUCTION

In this section I will discuss the inflationary model in the context of grand unified models of elementary-particle interactions.<sup>23,24</sup>

A grand unified model begins with a simple gauge group  $G$  which is a valid symmetry at the highest energies. As the energy is lowered, the theory undergoes a hierarchy of spontaneous symmetry breaking into successive subgroups:  $G \rightarrow H_n \rightarrow \dots \rightarrow H_0$ , where  $H_1 = \text{SU}_3 \times \text{SU}_2 \times U_1$  [QCD (quantum chromodynamics)  $\times$  Weinberg-Salam] and  $H_0 = \text{SU}_3 \times U_1^{\text{EM}}$ . In the Georgi-Glashow model,<sup>23</sup> which is the simplest model of this type,  $G = \text{SU}_5$  and  $n=1$ . The symmetry breaking of  $\text{SU}_5 \rightarrow \text{SU}_3 \times \text{SU}_2 \times U_1$  occurs at an energy scale  $M_X \sim 10^{14}$  GeV.

At high temperatures, it was suggested by Kirzhnits and Linde<sup>25</sup> that the Higgs fields of any spontaneously broken gauge theory would lose their expectation values, resulting in a high-temperature phase in which the full gauge symmetry is restored. A formalism for treating such problems was developed<sup>26</sup> by Weinberg and by Dolan and Jackiw. In the range of parameters for which the tree potential is valid, the phase structure of the  $\text{SU}_5$  model was analyzed by Tye and me.<sup>16,27</sup> We found that the  $\text{SU}_5$  symmetry is restored at  $T > \sim 10^{14}$  GeV and that for most values of the parameters there is an intermediate-temperature phase with gauge symmetry  $\text{SU}_4 \times U_1$ , which disappears at  $T \sim 10^{13}$  GeV. Thus, grand unified models tend to provide phase transitions which could lead to an inflationary scenario of the universe.

Grand unified models have another feature with important cosmological consequences: They contain very heavy magnetic monopoles in their particle spectrum. These monopoles are of the type discovered by 't Hooft and Polyakov,<sup>28</sup> and will be present in any model satisfying the above description.<sup>29</sup> These monopoles typically have masses of

order  $M_X/\alpha \sim 10^{16}$  GeV, where  $\alpha = g^2/4\pi$  is the grand unified fine structure constant. Since the monopoles are really topologically stable knots in the Higgs field expectation value, they do not exist in the high-temperature phase of the theory. They therefore come into existence during the course of a phase transition, and the dynamics of the phase transition is then intimately related to the monopole production rate.

The problem of monopole production and the subsequent annihilation of monopoles, in the context of a second-order or weakly first-order phase transition, was analyzed by Zeldovich and Khlopov<sup>30</sup> and by Preskill.<sup>31</sup> In Preskill's analysis, which was more specifically geared toward grand unified models, it was found that relic monopoles would exceed present bounds by roughly 14 orders of magnitude. Since it seems difficult to modify the estimated annihilation rate, one must find a scenario which suppresses the production of these monopoles.

Kibble<sup>32</sup> has pointed out that monopoles are produced in the course of the phase transition by the process of bubble coalescence. The orientation of the Higgs field inside one bubble will have no correlation with that of another bubble not in contact. When the bubbles coalesce to fill the space, it will be impossible for the uncorrelated Higgs fields to align uniformly. One expects to find topological knots, and these knots are the monopoles. The number of monopoles so produced is then comparable to the number of bubbles, to within a few orders of magnitude.

Kibble's production mechanism can be used to set a "horizon bound" on monopole production which is valid if the phase transition does not significantly disturb the evolution of the universe.<sup>33</sup> At the time of bubble coalescence  $t_{\text{coal}}$  the size  $l$  of the bubbles cannot exceed the horizon distance at that time. So

$$l < 2t_{\text{coal}} = \frac{M_P}{\gamma T_{\text{coal}}}^{\frac{1}{2}}. \quad (4.1)$$

By Kibble's argument, the density  $n_M$  of monopoles then obeys

$$n_M \geq l^{-3} > \frac{\gamma^3 T_{\text{coal}}^6}{M_P^3}. \quad (4.2)$$

By considering the contribution to the mass density of the present universe which could come from  $10^{16}$  GeV monopoles, Preskill<sup>31</sup> concludes that

$$n_M/n_\gamma < 10^{-24}, \quad (4.3)$$

where  $n_\gamma$  is the density of photons. This ratio changes very little from the time of the phase transition, so with (2.7) one concludes

$$T_{\text{coal}} < \left[ \frac{10^{-24} \pi^2}{2\xi(3)} \right]^{1/3} \gamma^{-1} M_P \approx 10^{10} \text{ GeV}. \quad (4.4)$$

If  $T_c \sim 10^{14}$  GeV, this bound implies that the universe must supercool by at least about four orders of magnitude before the phase transition is completed.

The problem of monopole production in a strongly first-order phase transition with supercooling was treated in more detail by Tye and me.<sup>16,34</sup> We showed how to explicitly calculate the bubble density in terms of the nucleation rate, and we considered the effects of the latent heat released in the phase transition. Our conclusion was that (4.4) should be replaced by

$$T_{\text{coal}} < 2 \times 10^{11} \text{ GeV}, \quad (4.5)$$

where  $T_{\text{coal}}$  refers to the temperature just before the release of the latent heat.

Tye and I omitted the crucial effects of the mass density  $\rho_0$  of the false vacuum. However, our work has one clear implication: If the nucleation rate is sufficiently large to avoid exponential growth, then far too many monopoles would be produced. Thus, the monopole problem seems to also force one into the inflationary scenario.<sup>35</sup>

In the simplest  $SU_5$  model, the nucleation rates have been calculated (approximately) by Weinberg and me.<sup>19</sup> The model contains unknown parameters, so no definitive answer is possible. We do find, however, that there is a sizable range of parameters which lead to the inflationary scenario.<sup>36</sup>

## V. PROBLEMS OF THE INFLATIONARY SCENARIO<sup>37</sup>

As I mentioned earlier, the inflationary scenario seems to lead to some unacceptable consequences. It is hoped that some variation can be found which avoids these undesirable features but maintains the desirable ones. The problems of the model will be discussed in more detail elsewhere,<sup>37</sup> but for completeness I will give a brief description here.

The central problem is the difficulty in finding a smooth ending to the period of exponential expansion. Let us assume that  $\lambda(t)$  approaches a constant as  $t \rightarrow \infty$  and  $T \rightarrow 0$ . To achieve the desired expansion factor  $Z > 10^{28}$ , one needs  $\lambda_0/\chi^4 < 10^{-2}$  [see (3.18)], which means that the nucleation rate is slow compared to the expansion rate of the universe. (Explicit calculations show that  $\lambda_0/\chi^4$  is typically much smaller than this value.<sup>18,19,36</sup>) The randomness of the bubble formation process then leads to gross inhomogeneities.

To understand the effects of this randomness, the reader should bear in mind the following facts.

(i) All of the latent heat released as a bubble expands is transferred initially to the walls of the

bubble.<sup>17</sup> This energy can be thermalized only when the bubble walls undergo many collisions.

(ii) The de Sitter metric does not single out a comoving frame. The  $O(4, 1)$  invariance of the de Sitter metric is maintained even after the formation of one bubble. The memory of the original Robertson-Walker comoving frame is maintained by the probability distribution of bubbles, but the local comoving frame can be reestablished only after enough bubbles have collided.

(iii) The size of the largest bubbles will exceed that of the smallest bubbles by roughly a factor of  $Z$ ; the range of bubble sizes is immense. The surface energy density grows with the size of the bubble, so the energy in the walls of the largest bubbles can be thermalized only by colliding with other large bubbles.

(iv) As time goes on, an arbitrarily large fraction of the space will be in the new phase [see (3.16)]. However, one can ask a more subtle question about the region of space which is in the new phase: Is the region composed of finite separated clusters, or do these clusters join together to form an infinite region? The latter possibility is called "percolation." It can be shown<sup>38</sup> that the system percolates for large values of  $\lambda_0/\chi^4$ , but that for sufficiently small values it does *not*. The critical value of  $\lambda_0/\chi^4$  has not been determined, but presumably an inflationary universe would have a value of  $\lambda_0/\chi^4$  below critical. Thus, no matter how long one waits, the region of space in the new phase will consist of finite clusters, each totally surrounded by a region in the old phase.

(v) Each cluster will contain only a few of the largest bubbles. Thus, the collisions discussed in (iii) cannot occur.

The above statements do not quite prove that the scenario is impossible, but these consequences are at best very unattractive. Thus, it seems that the scenario will become viable only if some modification can be found which avoids these inhomogeneities. Some vague possibilities will be mentioned in the next section.

Note that the above arguments seem to rule out the possibility that the universe was ever trapped in a false vacuum state, unless  $\lambda_0/\chi^4 \gtrsim 1$ . Such a large value of  $\lambda_0/\chi^4$  does not seem likely, but it is possible.<sup>19</sup>

## VI. CONCLUSION

I have tried to convince the reader that the standard model of the very early universe requires the assumption of initial conditions which are very implausible for two reasons:

(i) *The horizon problem.* Causally disconnected regions are assumed to be nearly identical; in par-

ticular, they are simultaneously at the same temperature.

(ii) *The flatness problem.* For a fixed initial temperature, the initial value of the Hubble "constant" must be fine tuned to extraordinary accuracy to produce a universe which is as flat as the one we observe.

Both of these problems would disappear if the universe supercooled by 28 or more orders of magnitude below the critical temperature for some phase transition. (Under such circumstances, the universe would be growing exponentially in time.) However, the random formation of bubbles of the new phase seems to lead to a much too inhomogeneous universe.

The inhomogeneity problem would be solved if one could avoid the assumption that the nucleation rate  $\lambda(t)$  approaches a small constant  $\lambda_0$  as the temperature  $T \rightarrow 0$ . If, instead, the nucleation rate rose sharply at some  $T_1$ , then bubbles of an approximately uniform size would suddenly fill space as  $T$  fell to  $T_1$ . Of course, the full advantage of the inflationary scenario is achieved only if  $T_1 \lesssim 10^{-28} T_c$ .

Recently Witten<sup>39</sup> has suggested that the above chain of events may in fact occur if the parameters of the  $SU_5$  Higgs field potential are chosen to obey the Coleman-Weinberg condition<sup>40</sup> (i.e., that  $\partial^2 V / \partial \phi^2 = 0$  at  $\phi = 0$ ). Witten<sup>41</sup> has studied this possibility in detail for the case of the Weinberg-Salam phase transition. Here he finds that thermal tunneling is totally ineffective, but instead the phase transition is driven when the temperature of the QCD chiral-symmetry-breaking phase transition is reached. For the  $SU_5$  case, one can hope that a much larger amount of supercooling will be found; however, it is difficult to see how 28 orders of magnitude could arise.

Another physical effect which has so far been left out of the analysis is the production of particles due to the changing gravitational metric.<sup>42</sup> This effect may become important in an exponentially expanding universe at low temperatures.

In conclusion, the inflationary scenario seems like a natural and simple way to eliminate both the horizon and the flatness problems. I am publishing this paper in the hope that it will highlight the existence of these problems and encourage others to find some way to avoid the undesirable features of the inflationary scenario.

#### ACKNOWLEDGMENTS

I would like to express my thanks for the advice and encouragement I received from Sidney Cole-

man and Leonard Susskind, and for the invaluable help I received from my collaborators Henry Tye and Erick Weinberg. I would also like to acknowledge very useful conversations with Michael Aizenman, Beilok Hu, Harry Kesten, Paul Langacker, Gordon Lasher, So-Young Pi, John Preskill, and Edward Witten. This work was supported by the Department of Energy under Contract No. DE-AC03-76SF00515.

#### APPENDIX: REMARKS ON THE FLATNESS PROBLEM

This appendix is added in the hope that some skeptics can be convinced that the flatness problem is real. Some physicists would rebut the argument given in Sec. I by insisting that the equations might make sense all the way back to  $t = 0$ . Then if one fixes the value of  $H$  corresponding to some arbitrary temperature  $T_a$ , one always finds that when the equations are extrapolated backward in time,  $\Omega \rightarrow 1$  as  $t \rightarrow 0$ . Thus, they would argue, it is natural for  $\Omega$  to be very nearly equal to 1 at early times. For physicists who take this point of view, the flatness problem must be restated in other terms. Since  $H_0$  and  $T_0$  have no significance, the model universe must be specified by its conserved quantities. In fact, the model universe is completely specified by the dimensionless constant  $\epsilon \equiv k/R^2 T^2$ , where  $k$  and  $R$  are parameters of the Robertson-Walker metric, Eq. (2.1). For our universe, one must take  $|\epsilon| < 3 \times 10^{-57}$ . The problem then is to explain why  $|\epsilon|$  should have such a startlingly small value.

Some physicists also take the point of view that  $\epsilon = 0$  is plausible enough, so to them there is no problem. To these physicists I point out that the universe is certainly not described *exactly* by a Robertson-Walker metric. Thus it is difficult to imagine any physical principle which would require a parameter of that metric to be exactly equal to zero.

In the end, I must admit that questions of plausibility are not logically determinable and depend somewhat on intuition. Thus I am sure that some physicists will remain unconvinced that there really is a flatness problem. However, I am also sure that many physicists agree with me that the flatness of the universe is a peculiar situation which at some point will admit a physical explanation.

\*Present address: Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

<sup>1</sup>I use units for which  $\hbar = c = k$  (Boltzmann constant) = 1. Then 1 m =  $5.068 \times 10^{15}$  GeV<sup>-1</sup>, 1 kg =  $5.610 \times 10^{26}$  GeV, 1 sec =  $1.519 \times 10^{24}$  GeV<sup>-1</sup>, and 1°K =  $8.617 \times 10^{-14}$  GeV.

<sup>2</sup>W. Rindler, Mon. Not. R. Astron. Soc. 116, 663 (1956). See also Ref. 3, pp. 489–490, 525–526; and Ref. 4, pp. 740 and 815.

<sup>3</sup>S. Weinberg, *Gravitation and Cosmology* (Wiley, New York, 1972).

<sup>4</sup>C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).

<sup>5</sup>In order to calculate the horizon distance, one must of course follow the light trajectories back to  $t = 0$ . This violates my contention that the equations are to be trusted only for  $T \lesssim T_0$ . Thus, the horizon problem could be obviated if the full quantum gravitational theory had a radically different behavior from the naive extrapolation. Indeed, solutions of this sort have been proposed by A. Zee, Phys. Rev. Lett. 44, 703 (1980) and by F. W. Stecker, Astrophys. J. 235, L1 (1980). However, it is the point of this paper to show that the horizon problem can also be obviated by mechanisms which are more within our grasp, occurring at temperatures below  $T_0$ .

<sup>6</sup>R. H. Dicke and P. J. E. Peebles, *General Relativity: An Einstein Centenary Survey*, edited by S. W. Hawking and W. Israel (Cambridge University Press, London, 1979).

<sup>7</sup>See Ref. 3, pp. 475–481; and Ref. 4, pp. 796–797.

<sup>8</sup>For example, see Ref. 3, Chap. 14.

<sup>9</sup>M. Yoshimura, Phys. Rev. Lett. 41, 281 (1978); 42, 746(E) (1979); Phys. Lett. 88B, 294 (1979); S. Dimopoulos and L. Susskind, Phys. Rev. D 18, 4500 (1978); Phys. Lett. 81B, 416 (1979); A. Yu Ignatiev, N. V. Krashikov, V. A. Kuzmin, and A. N. Tavkhelidze, *ibid.* 76B, 436 (1978); D. Toussaint, S. B. Treiman, F. Wilczek, and A. Zee, Phys. Rev. D 19, 1036 (1979); S. Weinberg, Phys. Rev. Lett. 42, 850 (1979); D. V. Nanopoulos and S. Weinberg, Phys. Rev. D 20, 2484 (1979); J. Ellis, M. K. Gaillard, and D. V. Nanopoulos, Phys. Lett. 80B, 360 (1979); 82B, 464 (1979); M. Honda and M. Yoshimura, Prog. Theor. Phys. 62, 1704 (1979); D. Toussaint and F. Wilczek, Phys. Lett. 81B, 238 (1979); S. Barr, G. Segre, and H. A. Weldon, Phys. Rev. D 20, 2494 (1979); A. D. Sakharov, Zh. Eksp. Teor. Fiz. 76, 1172 (1979) [Sov. Phys.—JETP 49, 594 (1979)]; A. Yu Ignatiev, N. V. Krashikov, V. A. Kuzmin, and M. E. Shaposhnikov, Phys. Lett. 87B, 114 (1979); E. W. Kolb and S. Wolfram, *ibid.* 91B, 217 (1980); Nucl. Phys. B172, 224 (1980); J. N. Fry, K. A. Olive, and M. S. Turner, Phys. Rev. D 22, 2953 (1980); 22, 2977 (1980); S. B. Treiman and F. Wilczek, Phys. Lett. 95B, 222 (1980); G. Senjanović and F. W. Stecker, Phys. Lett. B (to be published).

<sup>11</sup>The reason  $\Lambda$  is so small is of course one of the deep mysteries of physics. The value of  $\Lambda$  is not determined by the particle theory alone, but must be fixed by whatever theory couples particles to quantum gravity. This appears to be a separate problem from the ones discussed in this paper, and I merely use the empirical fact that  $\Lambda \approx 0$ .

<sup>12</sup>S. A. Bludman and M. A. Ruderman, Phys. Rev. Lett.

38, 255 (1977).

<sup>13</sup>The effects of a false vacuum energy density on the evolution of the early universe have also been considered by E. W. Kolb and S. Wolfram, CAL TECH Report No. 79-0984 (unpublished), and by S. A. Bludman, University of Pennsylvania Report No. UPR-0143T, 1979 (unpublished).

<sup>14</sup>More precisely, the usual invariance is O(4) if  $k = 1$ , O(3,1) if  $k = -1$ , and the group of rotations and translations in three dimensions if  $k = 0$ .

<sup>15</sup>See for example, Ref. 3, pp. 385–392.

<sup>16</sup>A. H. Guth and S.-H. Tye, Phys. Rev. Lett. 44, 631 (1980); 44, 963 (1980).

<sup>17</sup>S. Coleman, Phys. Rev. D 15, 2929 (1977); C. G. Callan and S. Coleman, *ibid.* 16, 1762 (1977); see also S. Coleman, in *The Whys of Subnuclear Physics*, proceedings of the International School of Subnuclear Physics, Ettore Majorana, Erice, 1977, edited by A. Zichichi (Plenum, New York, 1979).

<sup>18</sup>A. H. Guth and E. J. Weinberg, Phys. Rev. Lett. 45, 1131 (1980).

<sup>19</sup>E. J. Weinberg and I are preparing a manuscript on the possible cosmological implications of the phase transitions in the SU<sub>5</sub> grand unified model.

<sup>20</sup>J. Ellis and G. Steigman, Phys. Lett. 89B, 186 (1980); J. Ellis, M. K. Gaillard, and D. V. Nanopoulos, *ibid.* 90B, 253 (1980).

<sup>21</sup>B. L. Hu and L. Parker, Phys. Rev. D 17, 933 (1978).

<sup>22</sup>See Ref. 4, Chap. 30.

<sup>23</sup>The simplest grand unified model is the SU(5) model of H. Georgi and S. L. Glashow, Phys. Rev. Lett. 32, 438 (1974). See also H. Georgi, H. R. Quinn, and S. Weinberg, *ibid.* 33, 451 (1974); and A. J. Buras, J. Ellis, M. K. Gaillard, and D. V. Nanopoulos, Nucl. Phys. B135, 66 (1978).

<sup>24</sup>Other grand unified models include the SO(10) model: H. Georgi, in *Particles and Fields—1975*, proceedings of the 1975 meeting of the Division of Particles and Fields of the American Physical Society, edited by Carl Carlson (AIP, New York, 1975); H. Fritzsch and P. Minkowski, Ann. Phys. (N.Y.) 93, 193 (1975); H. Georgi and D. V. Nanopoulos, Phys. Lett. 82B, 392 (1979) and Nucl. Phys. B155, 52 (1979). The E(6) model: F. Gürsey, P. Ramond, and P. Sikivie, Phys. Lett. 60B, 177 (1975); F. Gürsey and M. Serdaroglu, Lett. Nuovo Cimento 21, 28 (1978). The E(7) model: F. Gürsey and P. Sikivie, Phys. Rev. Lett. 36, 775 (1976), and Phys. Rev. D 16, 816 (1977); P. Ramond, Nucl. Phys. B110, 214 (1976). For some general properties of grand unified models, see M. Gell-Mann, P. Ramond, and R. Slansky, Rev. Mod. Phys. 50, 721 (1978). For a review, see P. Langacker, Report No. SLAC-PUB-2544, 1980 (unpublished).

<sup>25</sup>D. A. Kirzhnits and A. D. Linde, Phys. Lett. 42B, 471 (1972).

<sup>26</sup>S. Weinberg, Phys. Rev. D 9, 3357 (1974); L. Dolan and R. Jackiw, *ibid.* 9, 3320 (1974); see also D. A. Kirzhnits and A. D. Linde, Ann. Phys. (N.Y.) 101, 195 (1976); A. D. Linde, Rep. Prog. Phys. 42, 389 (1979).  $\epsilon$ -expansion techniques are employed by P. Ginsparg, Nucl. Phys. B (to be published).

<sup>27</sup>In the case that the Higgs quartic couplings are comparable to  $g^4$  or smaller ( $g$  = gauge coupling), the phase structure has been studied by M. Daniel and C. E. Vayonakis, CERN Report No. TH.2860 1980

- (unpublished); and by P. Suranyi, University of Cincinnati Report No. 80-0506 (unpublished).
- <sup>28</sup>G. 't Hooft, Nucl. Phys. B79, 276 (1974); A. M. Polyakov, Pis'ma Zh. Eksp. Teor. Fiz. 20, 430 (1974) [JETP Lett. 20, 194 (1974)]. For a review, see P. Goddard and D. I. Olive, Rep. Prog. Phys. 41, 1357 (1978).
- <sup>29</sup>If  $\Pi_1(G)$  and  $\Pi_2(G)$  are both trivial, then  $\Pi_2(G/H_0) = \Pi_1(H_0)$ . In our case  $\Pi_1(H_0)$  is the group of integers. For a general review of topology written for physicists, see N. D. Mermin, Rev. Mod. Phys. 51, 591 (1979).
- <sup>30</sup>Y. B. Zeldovich and M. Y. Khlopov, Phys. Lett. 79B, 239 (1978).
- <sup>31</sup>J. P. Preskill, Phys. Rev. Lett. 43, 1365 (1979).
- <sup>32</sup>T. W. B. Kibble, J. Phys. A 9, 1387 (1976).
- <sup>33</sup>This argument was first shown to me by John Preskill. It is also described by Einhorn *et al.*, Ref. 34, except that they make no distinction between  $T_{\text{coal}}$  and  $T_c$ .
- <sup>34</sup>The problem of monopole production was also examined by M. B. Einhorn, D. L. Stein, and D. Toussaint, Phys. Rev. D 21, 3295 (1980), who focused on second-order transitions. The structure of  $SU(5)$  monopoles has been studied by C. P. Dokos and T. N. Tomaras, Phys. Rev. D 21, 2940 (1980); and by M. Daniel, G. Lazarides, and Q. Shafi, Nucl. Phys. B170, 156 (1980). The problem of suppression of the cosmological production of monopoles is discussed by G. Lazarides and Q. Shafi, Phys. Lett. 94B, 149 (1980), and G. Lazarides, M. Magg, and Q. Shafi, CERN Report No. TH.2856, 1980 (unpublished); the suppression discussed here relies on a novel confinement mechanism, and also on the same kind of supercooling as in Ref. 16. See also J. N. Fry and D. N. Schramm, Phys. Rev. Lett. 44, 1361 (1980).
- <sup>35</sup>An alternative solution to the monopole problem has been proposed by P. Langacker and S.-Y. Pi, Phys. Rev. Lett. 45, 1 (1980). By modifying the Higgs structure, they have constructed a model in which the high-temperature  $SU_5$  phase undergoes a phase transition to an  $SU_3$  phase at  $T \sim 10^{14}$  GeV. Another phase transition occurs at  $T \sim 10^3$  GeV, and below this temperature the symmetry is  $SU_3 \times U_1^{\text{EM}}$ . Monopoles cannot exist until  $T < 10^3$  GeV, but their production is negligible at these low temperatures. The suppression of monopoles due to the breaking of  $U_1^{\text{EM}}$  symmetry at high temperatures was also suggested by S. -H. Tye, talk given at the 1980 Guangzhou Conference on Theoretical Particle Physics, Canton, 1980 (unpublished).
- <sup>36</sup>The Weinberg-Salam phase transition has also been investigated by a number of authors: E. Witten, Ref. 41; M. A. Sher, Phys. Rev. D 22, 2989 (1980); P. J. Steinhardt, Harvard report, 1980 (unpublished); and A. H. Guth and E. J. Weinberg, Ref. 18.
- <sup>37</sup>This section represents the work of E. J. Weinberg, H. Kesten, and myself. Weinberg and I are preparing a manuscript on this subject.
- <sup>38</sup>The proof of this statement was outlined by H. Kesten (Dept. of Mathematics, Cornell University), with details completed by me.
- <sup>39</sup>E. Witten, private communication.
- <sup>40</sup>S. Coleman and E. J. Weinberg, Phys. Rev. D 7, 1888 (1973); see also, J. Ellis, M. K. Gaillard, D. Nanopoulos, and C. Sachrajda, Phys. Lett. 83B, 339 (1979), and J. Ellis, M. K. Gaillard, A. Peterman, and C. Sachrajda, Nucl. Phys. B164, 253 (1980).
- <sup>41</sup>E. Witten, Nucl. Phys. B (to be published).
- <sup>42</sup>L. Parker, in *Asymptotic Structure of Spacetime*, edited by F. Esposito and L. Witten (Plenum, New York, 1977); V. N. Lukash, I. D. Novikov, A. A. Starobinsky, and Ya. B. Zeldovich, Nuovo Cimento 35B, 293 (1976).

## **Simulating Physics with Computers**

**Richard P. Feynman**

*Department of Physics, California Institute of Technology, Pasadena, California 91107*

*Received May 7, 1981*

### **1. INTRODUCTION**

On the program it says this is a keynote speech—and I don't know what a keynote speech is. I do not intend in any way to suggest what should be in this meeting as a keynote of the subjects or anything like that. I have my own things to say and to talk about and there's no implication that anybody needs to talk about the same thing or anything like it. So what I want to talk about is what Mike Dertouzos suggested that nobody would talk about. I want to talk about the problem of simulating physics with computers and I mean that in a specific way which I am going to explain. The reason for doing this is something that I learned about from Ed Fredkin, and my entire interest in the subject has been inspired by him. It has to do with learning something about the possibilities of computers, and also something about possibilities in physics. If we suppose that we know all the physical laws perfectly, of course we don't have to pay any attention to computers. It's interesting anyway to entertain oneself with the idea that we've got something to learn about physical laws; and if I take a relaxed view here (after all I'm here and not at home) I'll admit that we don't understand everything.

The first question is, What kind of computer are we going to use to simulate physics? Computer theory has been developed to a point where it realizes that it doesn't make any difference; when you get to a *universal computer*, it doesn't matter how it's manufactured, how it's actually made. Therefore my question is, Can physics be simulated by a universal computer? I would like to have the elements of this computer *locally interconnected*, and therefore sort of think about cellular automata as an example (but I don't want to force it). But I do want something involved with the

locality of interaction. I would not like to think of a very enormous computer with arbitrary interconnections throughout the entire thing.

Now, what kind of physics are we going to imitate? First, I am going to describe the possibility of simulating physics in the classical approximation, a thing which is usually described by local differential equations. But the physical world is quantum mechanical, and therefore the proper problem is the simulation of quantum physics—which is what I really want to talk about, but I'll come to that later. So what kind of simulation do I mean? There is, of course, a kind of approximate simulation in which you design numerical algorithms for differential equations, and then use the computer to compute these algorithms and get an approximate view of what physics ought to do. That's an interesting subject, but is not what I want to talk about. I want to talk about the possibility that there is to be an *exact* simulation, that the computer will do *exactly* the same as nature. If this is to be proved and the type of computer is as I've already explained, then it's going to be necessary that *everything* that happens in a finite volume of space and time would have to be exactly analyzable with a finite number of logical operations. The present theory of physics is not that way, apparently. It allows space to go down into infinitesimal distances, wavelengths to get infinitely great, terms to be summed in infinite order, and so forth; and therefore, if this proposition is right, physical law is wrong.

So good, we already have a suggestion of how we might modify physical law, and that is the kind of reason why I like to study this sort of problem. To take an example, we might change the idea that space is continuous to the idea that space perhaps is a simple lattice and everything is discrete (so that we can put it into a finite number of digits) and that time jumps discontinuously. Now let's see what kind of a physical world it would be or what kind of problem of computation we would have. For example, the first difficulty that would come out is that the speed of light would depend slightly on the direction, and there might be other anisotropies in the physics that we could detect experimentally. They might be very small anisotropies. Physical knowledge is of course always incomplete, and you can always say we'll try to design something which beats experiment at the present time, but which predicts anisotropies on some scale to be found later. That's fine. That would be good physics if you could predict something consistent with all the known facts and suggest some new fact that we didn't explain, but I have no specific examples. So I'm not objecting to the fact that it's anisotropic in principle, it's a question of how anisotropic. If you tell me it's so-and-so anisotropic, I'll tell you about the experiment with the lithium atom which shows that the anisotropy is less than that much, and that this here theory of yours is impossible.

Another thing that had been suggested early was that natural laws are reversible, but that computer rules are not. But this turned out to be false; the computer rules can be reversible, and it has been a very, very useful thing to notice and to discover that. (Editors' note: see papers by Bennett, Fredkin, and Toffoli, these Proceedings). This is a place where the relationship of physics and computation has turned itself the other way and told us something about the possibilities of computation. So this is an interesting subject because it tells us something about computer rules, and *might* tell us something about physics.

The rule of simulation that I would like to have is that the number of computer elements required to simulate a large physical system is only to be proportional to the space-time volume of the physical system. I don't want to have an explosion. That is, if you say I want to explain this much physics, I can do it exactly and I need a certain-sized computer. If doubling the volume of space and time means I'll need an *exponentially* larger computer, I consider that against the rules (I make up the rules, I'm allowed to do that). Let's start with a few interesting questions.

## 2. SIMULATING TIME

First I'd like to talk about simulating time. We're going to assume it's discrete. You know that we don't have infinite accuracy in physical measurements so time might be discrete on a scale of less than  $10^{-27}$  sec. (You'd have to have it at least like this to avoid clashes with experiment—but make it  $10^{-41}$  sec. if you like, and then you've got us!)

One way in which we simulate time—in cellular automata, for example—is to say that “the computer goes from state to state.” But really, that's using intuition that involves the idea of time—you're going from state to state. And therefore the time (by the way, like the space in the case of cellular automata) is not simulated at all, it's imitated in the computer.

An interesting question comes up: “Is there a way of simulating it, rather than imitating it?” Well, there's a way of looking at the world that is called the space-time view, imagining that the points of space and time are all laid out, so to speak, ahead of time. And then we could say that a “computer” rule (now computer would be in quotes, because it's not the standard kind of computer which operates in time) is: We have a state  $s_i$  at each point  $i$  in space-time. (See Figure 1.) The state  $s_i$  at the space time point  $i$  is a given function  $F_i(s_j, s_k, \dots)$  of the state at the points  $j, k$  in some neighborhood of  $i$ :

$$s_i = F_i(s_j, s_k, \dots)$$

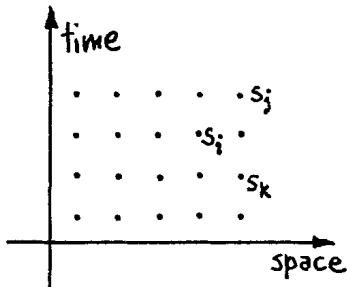


Fig. 1.

You'll notice immediately that if this particular function is such that the value of the function at  $i$  only involves the few points behind in time, earlier than this time  $i$ , all I've done is to redescribe the cellular automaton, because it means that you calculate a given point from points at earlier times, and I can compute the next one and so on, and I can go through this in that particular order. But just let's us think of a more general kind of computer, because we might have a more general function. So let's think about whether we could have a wider case of generality of interconnections of points in space-time. If  $F$  depends on *all* the points both in the future and the past, what then? That could be the way physics works. I'll mention how our theories go at the moment. It has turned out in many physical theories that the mathematical equations are quite a bit simplified by imagining such a thing—by imagining positrons as electrons going backwards in time, and other things that connect objects forward and backward. The important question would be, if this computer were laid out, is there in fact an organized algorithm by which a solution could be laid out, that is, computed? Suppose you know this function  $F_i$  and it is a function of the variables in the future as well. How would you lay out numbers so that they automatically satisfy the above equation? It may not be possible. In the case of the cellular automaton it is, because from a given row you get the next row and then the next row, and there's an organized way of doing it. It's an interesting question whether there are circumstances where you get functions for which you can't think, at least right away, of an organized way of laying it out. Maybe sort of shake it down from some approximation, or something, but it's an interesting different type of computation.

Question: "Doesn't this reduce to the ordinary boundary value, as opposed to initial-value type of calculation?"

Answer: "Yes, but remember this is the computer itself that I'm describing."

It appears actually that classical physics is causal. You can, in terms of the information in the past, if you include both momentum and position, or

the position at two different times in the past (either way, you need two pieces of information at each point) calculate the future in principle. So classical physics is *local*, *causal*, and *reversible*, and therefore apparently quite adaptable (except for the discreteness and so on, which I already mentioned) to computer simulation. We have no difficulty, in principle, apparently, with that.

### 3. SIMULATING PROBABILITY

Turning to quantum mechanics, we know immediately that here we get only the ability, apparently, to predict probabilities. Might I say immediately, so that you know where I really intend to go, that we always have had (secret, secret, close the doors!) we always have had a great deal of difficulty in understanding the world view that quantum mechanics represents. At least I do, because I'm an old enough man that I haven't got to the point that this stuff is obvious to me. Okay, I still get nervous with it. And therefore, some of the younger students ... you know how it always is, every new idea, it takes a generation or two until it becomes obvious that there's no real problem. It has not yet become obvious to me that there's no real problem. I cannot define the real problem, therefore I suspect there's no real problem, but I'm not sure there's no real problem. So that's why I like to investigate things. Can I learn anything from asking this question about computers—about this may or may not be mystery as to what the world view of quantum mechanics is? So I know that quantum mechanics seem to involve probability—and I therefore want to talk about simulating probability.

Well, one way that we could have a computer that simulates a probabilistic theory, something that has a probability in it, would be to calculate the probability and then interpret this number to represent nature. For example, let's suppose that a particle has a probability  $P(x, t)$  to be at  $x$  at a time  $t$ . A typical example of such a probability might satisfy a differential equation, as, for example, if the particle is diffusing:

$$\frac{\partial P(x, t)}{\partial t} = -\nabla^2 P(x, t)$$

Now we could discretize  $t$  and  $x$  and perhaps even the probability itself and solve this differential equation like we solve any old field equation, and make an algorithm for it, making it exact by discretization. First there'd be a problem about discretizing probability. If you are only going to take  $k$  digits it would mean that when the probability is less than  $2^{-k}$  of something happening, you say it doesn't happen at all. In practice we do that. If the

probability of something is  $10^{-700}$ , we say it isn't going to happen, and we're not caught out very often. So we could allow ourselves to do that. But the real difficulty is this: If we had many particles, we have  $R$  particles, for example, in a system, then we would have to describe the probability of a circumstance by giving the probability to find these particles at points  $x_1, x_2, \dots, x_R$  at the time  $t$ . That would be a description of the probability of the system. And therefore, you'd need a  $k$ -digit number for every configuration of the system, for every arrangement of the  $R$  values of  $x$ . And therefore if there are  $N$  points in space, we'd need  $N^R$  configurations. Actually, from our point of view that at each point in space there is information like electric fields and so on,  $R$  will be of the same order as  $N$  if the number of information bits is the same as the number of points in space, and therefore you'd have to have something like  $N^N$  configurations to be described to get the probability out, and that's too big for our computer to hold if the size of the computer is of order  $N$ .

We emphasize, if a description of an isolated part of nature with  $N$  variables requires a general function of  $N$  variables and if a computer stimulates this by actually computing or storing this function then doubling the size of nature ( $N \rightarrow 2N$ ) would require an exponentially explosive growth in the size of the simulating computer. It is therefore impossible, according to the rules stated, to simulate by calculating the probability.

Is there any other way? What kind of simulation can we have? We can't expect to compute the probability of configurations for a probabilistic theory. But the other way to simulate a probabilistic nature, which I'll call  $\mathcal{N}$  for the moment, might still be to simulate the probabilistic nature by a computer  $\mathcal{C}$  which itself is probabilistic, in which you always randomize the last two digits of every number, or you do something terrible to it. So it becomes what I'll call a probabilistic computer, in which the output is not a unique function of the input. And then you try to work it out so that it simulates nature in this sense: that  $\mathcal{C}$  goes from some state—initial state if you like—to some final state with the *same* probability that  $\mathcal{N}$  goes from the corresponding initial state to the corresponding final state. Of course when you set up the machine and let nature do it, the imitator will not do the same thing, it only does it with the same probability. Is that no good? No it's O.K. How do you know what the probability is? You see, nature's unpredictable; how do you expect to predict it with a computer? You can't, —it's unpredictable if it's probabilistic. But what you really do in a probabilistic system is repeat the experiment in nature a large number of times. If you repeat the same experiment in the computer a large number of times (and that doesn't take any more time than it does to do the same thing in nature of course), it will give the frequency of a given final state proportional to the number of times, with approximately the same rate (plus

or minus the square root of  $n$  and all that) as it happens in nature. In other words, we could imagine and be perfectly happy, I think, with a probabilistic simulator of a probabilistic nature, in which the machine doesn't exactly do what nature does, but if you repeated a particular type of experiment a sufficient number of times to determine nature's probability, then you did the corresponding experiment on the computer, you'd get the corresponding probability with the corresponding accuracy (with the same kind of accuracy of statistics).

So let us now think about the characteristics of a local probabilistic computer, because I'll see if I can imitate nature with that (by "nature" I'm now going to mean quantum mechanics). One of the characteristics is that you can determine how it behaves in a local region by simply disregarding what it's doing in all other regions. For example, suppose there are variables in the system that describe the whole world ( $x_A, x_B$ )—the variables  $x_A$  you're interested in, they're "around here";  $x_B$  are the whole result of the world. If you want to know the probability that something around here is happening, you would have to get that by integrating the total probability of all kinds of possibilities over  $x_B$ . If we had *computed* this probability, we would still have to do the integration

$$P_A(x_A) = \int P(x_A, x_B) dx_B$$

which is a hard job! But if we have *imitated* the probability, it's very simple to do it: you don't have to do anything to do the integration, you simply disregard what the values of  $x_B$  are, you just look at the region  $x_A$ . And therefore it does have the characteristic of nature: if it's local, you can find out what's happening in a region not by integrating or doing an extra operation, but merely by disregarding what happens elsewhere, which is no operation, nothing at all.

The other aspect that I want to emphasize is that the equations will have a form, no doubt, something like the following. Let each point  $i = 1, 2, \dots, N$  in space be in a state  $s_i$  chosen from a small state set (the size of this set should be reasonable, say, up to  $2^5$ ). And let the probability to find some configuration  $\{s_i\}$  (a set of values of the state  $s_i$  at each point  $i$ ) be some number  $P(\{s_i\})$ . It satisfies an equation such that at each jump in time

$$P_{t+1}(\{s\}) = \sum_{\{s'\}} \left[ \prod_i m(s_i | s'_j, s'_{k\dots}) \right] P_t(\{s'\})$$

where  $m(s_i | s'_j, s'_{k\dots})$  is the probability that we move to state  $s_i$  at point  $i$

when the neighbors have values  $s'_j, s'_k\dots$ , where  $j, k$  etc. are points in the neighborhood of  $i$ . As  $j$  moves far from  $i$ ,  $m$  becomes ever less sensitive to  $s'_j$ . At each change the state at a particular point  $i$  will move from what it was to a state  $s$  with a probability  $m$  that depends only upon the states of the neighborhood (which may be so defined as to include the point  $i$  itself). This gives the probability of making a transition. It's the same as in a cellular automaton; only, instead of its being definite, it's a probability. Tell me the environment, and I'll tell you the probability after a next moment of time that this point is at state  $s$ . And that's the way it's going to work, okay? So you get a mathematical equation of this kind of form.

Now I explicitly go to the question of how we can simulate with a computer—a universal automaton or something—the quantum-mechanical effects. (The usual formulation is that quantum mechanics has some sort of a differential equation for a function  $\psi$ .) If you have a single particle,  $\psi$  is a function of  $x$  and  $t$ , and this differential equation could be simulated just like my probabilistic equation was before. That would be all right and one has seen people make little computers which simulate the Schrödinger equation for a single particle. But the full description of quantum mechanics for a large system with  $R$  particles is given by a function  $\psi(x_1, x_2, \dots, x_R, t)$  which we call the amplitude to find the particles  $x_1, \dots, x_R$ , and therefore, because it has too many variables, it *cannot be simulated* with a normal computer with a number of elements proportional to  $R$  or proportional to  $N$ . We had the same troubles with the probability in classical physics. And therefore, the problem is, how can we simulate the quantum mechanics? There are two ways that we can go about it. We can give up on our rule about what the computer was, we can say: Let the computer itself be built of quantum mechanical elements which obey quantum mechanical laws. Or we can turn the other way and say: Let the computer still be the same kind that we thought of before—a logical, universal automaton; can we imitate this situation? And I'm going to separate my talk here, for it branches into two parts.

#### 4. QUANTUM COMPUTERS—UNIVERSAL QUANTUM SIMULATORS

The first branch, one you might call a side-remark, is, Can you do it with a new kind of computer—a quantum computer? (I'll come back to the other branch in a moment.) Now it turns out, as far as I can tell, that you can simulate this with a quantum system, with quantum computer elements. It's not a Turing machine, but a machine of a different kind. If we disregard the continuity of space and make it discrete, and so on, as an approximation (the same way as we allowed ourselves in the classical case), it does seem to

be true that all the various field theories have the same *kind* of behavior, and can be simulated in every way, apparently, with little latticeworks of spins and other things. It's been noted time and time again that the phenomena of field theory (if the world is made in a discrete lattice) are well imitated by many phenomena in solid state theory (which is simply the analysis of a latticework of crystal atoms, and in the case of the kind of solid state I mean each atom is just a point which has numbers associated with it, with quantum-mechanical rules). For example, the spin waves in a spin lattice imitating Bose-particles in the field theory. I therefore believe it's true that with a suitable class of quantum machines you could imitate any quantum system, including the physical world. But I don't know whether the general theory of this intersimulation of quantum systems has ever been worked out, and so I present that as another interesting problem: to work out the classes of different kinds of quantum mechanical systems which are really intersimulatable—which are equivalent—as has been done in the case of classical computers. It has been found that there is a kind of universal computer that can do anything, and it doesn't make much difference specifically how it's designed. The same way we should try to find out what kinds of quantum mechanical systems are mutually intersimulatable, and try to find a specific class, or a character of that class which will simulate everything. What, in other words, is the universal quantum simulator? (assuming this discretization of space and time). If you had discrete quantum systems, what other discrete quantum systems are exact imitators of it, and is there a class against which everything can be matched? I believe it's rather simple to answer that question and to find the class, but I just haven't done it.

Suppose that we try the following guess: that every finite quantum mechanical system can be described *exactly*, imitated exactly, by supposing that we have another system such that at each point in space-time this system has only two possible base states. Either that point is occupied, or unoccupied—those are the two states. The mathematics of the quantum mechanical operators associated with that point would be very simple.

$$\begin{array}{l}
 a = \text{ANNIHILATE} = \begin{array}{c|cc} & \text{OCC} & \text{UN} \\ \hline \text{OCC} & 0 & 0 \\ \text{UN} & 1 & 0 \end{array} = \frac{1}{2}(\sigma_x - i\sigma_y) \\
 a^* = \text{CREATE} = \begin{array}{c|cc} & 0 & 1 \\ \hline 0 & 0 & 0 \end{array} = \frac{1}{2}(\sigma_x + i\sigma_y) \\
 n = \text{NUMBER} = \begin{array}{c|cc} & 1 & 0 \\ \hline 0 & 0 & 0 \end{array} = a^*a = \frac{1}{2}(1 + \sigma_z) \\
 1 = \text{IDENTITY} = \begin{array}{c|cc} & 1 & 0 \\ \hline 0 & 0 & 1 \end{array}
 \end{array}$$

There would be an operator  $a$  which *annihilates* if the point is occupied—it changes it to unoccupied. There is a conjugate operator  $a^*$  which does the opposite: if it's unoccupied, it occupies it. There's another operator  $n$  called the *number* to ask, Is something there? The little matrices tell you what they do. If it's there,  $n$  gets a one and leaves it alone, if it's not there, nothing happens. That's mathematically equivalent to the product of the other two, as a matter of fact. And then there's the identity,  $1$ , which we always have to put in there to complete our mathematics—it doesn't do a damn thing!

By the way, on the right-hand side of the above formulas the same operators are written in terms of matrices that most physicists find more convenient, because they are Hermitian, and that seems to make it easier for them. They have invented another set of matrices, the Pauli  $\sigma$  matrices:

$$\sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad 1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

And these are called *spin*—spin one-half—so sometimes people say you're talking about a spin-one-half lattice.

The question is, if we wrote a Hamiltonian which involved only these operators, locally coupled to corresponding operators on the other space-time points, could we imitate every quantum mechanical system which is discrete and has a finite number of degrees of freedom? I know, almost certainly, that we could do that for any quantum mechanical system which involves Bose particles. I'm not sure whether Fermi particles could be described by such a system. So I leave that open. Well, that's an example of what I meant by a general quantum mechanical simulator. I'm not sure that it's sufficient, because I'm not sure that it takes care of Fermi particles.

## 5. CAN QUANTUM SYSTEMS BE PROBABILISTICALLY SIMULATED BY A CLASSICAL COMPUTER?

Now the next question that I would like to bring up is, of course, the interesting one, i.e., Can a quantum system be probabilistically simulated by a classical (probabilistic, I'd assume) universal computer? In other words, a computer which will give the same probabilities as the quantum system does. If you take the computer to be the classical kind I've described so far, (not the quantum kind described in the last section) and there're no changes in any laws, and there's no hocus-pocus, the answer is certainly, No! This is called the hidden-variable problem: it is impossible to represent the results of quantum mechanics with a classical universal device. To learn a little bit about it, I say let us try to put the quantum equations in a form as close as

possible to classical equations so that we can see what the difficulty is and what happens. Well, first of all we can't simulate  $\psi$  in the normal way. As I've explained already, there're too many variables. Our only hope is that we're going to simulate probabilities, that we're going to have our computer do things with the same probability as we observe in nature, as calculated by the quantum mechanical system. Can you make a cellular automaton, or something, imitate with the same probability what nature does, where I'm going to suppose that quantum mechanics is correct, or at least after I discretize space and time it's correct, and see if I can do it. I must point out that you must directly generate the probabilities, the results, with the correct quantum probability. Directly, because we have no way to store all the numbers, we have to just imitate the phenomenon directly.

It turns out then that another thing, rather than the wave function, a thing called the *density matrix*, is much more useful for this. It's not so useful as far as the mathematical equations are concerned, since it's more complicated than the equations for  $\psi$ , but I'm not going to worry about mathematical complications, or which is the easiest way to calculate, because with computers we don't have to be so careful to do it the very easiest way. And so with a slight increase in the complexity of the equations (and not very much increase) I turn to the density matrix, which for a single particle of coordinate  $x$  in a pure state of wave function  $\psi(x)$  is

$$\rho(x, x') = \psi^*(x)\psi(x')$$

This has a special property that is a function of two coordinates  $x, x'$ . The presence of two quantities  $x$  and  $x'$  associated with each coordinate is analogous to the fact that in classical mechanics you have to have two variables to describe the state,  $x$  and  $\dot{x}$ . States are described by a second-order device, with two informations ("position" and "velocity"). So we have to have two pieces of information associated with a particle, analogous to the classical situation, in order to describe configurations. (I've written the density matrix for one particle, but of course there's the analogous thing for  $R$  particles, a function of  $2R$  variables).

This quantity has many of the mathematical properties of a probability. For example if a state  $\psi(x)$  is not certain but is  $\psi_\alpha$  with the probability  $p_\alpha$  then the density matrix is the appropriate weighted sum of the matrix for each state  $\alpha$ :

$$\rho(x, x') = \sum_{\alpha} p_{\alpha} \psi_{\alpha}^*(x) \psi_{\alpha}(x').$$

A quantity which has properties even more similar to classical probabilities is the Wigner function, a simple reexpression of the density matrix; for a

single particle

$$W(x, p) = \int \rho\left(x + \frac{y}{2}, x - \frac{y}{2}\right) e^{ipy} dy$$

We shall be emphasizing their similarity and shall call it “probability” in quotes instead of Wigner function. Watch these quotes carefully, when they are absent we mean the real probability. If “probability” had all the mathematical properties of a probability we could remove the quotes and simulate it.  $W(x, p)$  is the “probability” that the particle has position  $x$  and momentum  $p$  (per  $dx$  and  $dp$ ). What properties does it have that are analogous to an ordinary probability?

It has the property that if there are many variables and you want to know the “probabilities” associated with a finite region, you simply disregard the other variables (by integration). Furthermore the probability of finding a particle at  $x$  is  $\int W(x, p) dp$ . If you can interpret  $W$  as a probability of finding  $x$  and  $p$ , this would be an expected equation. Likewise the probability of  $p$  would be expected to be  $\int W(x, p) dx$ . These two equations are correct, and therefore you would hope that maybe  $W(x, p)$  is the probability of finding  $x$  and  $p$ . And the question then is can we make a device which simulates this  $W$ ? Because then it would work fine.

Since the quantum systems I noted were best represented by spin one-half (occupied versus unoccupied or spin one-half is the same thing), I tried to do the same thing for spin one-half objects, and it's rather easy to do. Although before one object only had two states, occupied and unoccupied, the full description—in order to develop things as a function of time—requires twice as many variables, which mean two slots at each point which are occupied or unoccupied (denoted by + and - in what follows), analogous to the  $x$  and  $\dot{x}$ , or the  $x$  and  $p$ . So you can find four numbers, four “probabilities”  $\{f_{++}, f_{+-}, f_{-+}, f_{--}\}$  which act just like, and I have to explain why they're not exactly like, but they act just like, probabilities to find things in the state in which both symbols are up, one's up and one's down, and so on. For example, the sum  $f_{++} + f_{+-} + f_{-+} + f_{--}$  of the four “probabilities” is 1. You'll remember that one object now is going to have two indices, two plus/minus indices, or two ones and zeros at each point, although the quantum system had only one. For example, if you would like to know whether the first index is positive, the probability of that would be

$$\text{Prob(first index is +)} = f_{++} + f_{+-} \quad [\text{spin } z \text{ up}]$$

i.e., you don't care about the second index. The probability that the first index is negative is

$$\text{Prob(first index is -)} = f_{-+} + f_{--} \quad [\text{spin } z \text{ down}]$$

These two formulas are exactly correct in quantum mechanics. You see I'm hedging on whether or not "probability"  $f$  can really be a probability without quotes. But when I write probability without quotes on the left-hand side I'm not hedging; that really is the quantum mechanical probability. It's interpreted perfectly fine here. Likewise the probability that the second index is positive can be obtained by finding

$$\text{Prob(second index is +)} = f_{++} + f_{-+} \quad [\text{spin } x \text{ up}]$$

and likewise

$$\text{Prob(second index is -)} = f_{+-} + f_{--} \quad [\text{spin } x \text{ down}]$$

You could also ask other questions about the system. You might like to know, What is the probability that both indices are positive? You'll get in trouble. But you could ask other questions that you won't get in trouble with, and that get correct physical answers. You can ask, for example, what is the probability that the two indices are the same? That would be

$$\text{Prob(match)} = f_{++} + f_{--} \quad [\text{spin } y \text{ up}]$$

Or the probability that there's no match between the indices, that they're different,

$$\text{Prob(no match)} = f_{+-} + f_{-+} \quad [\text{spin } y \text{ down}]$$

All perfectly all right. All these probabilities are correct and make sense, and have a precise meaning in the spin model, shown in the square brackets above. There are other "probability" combinations, other linear combinations of these  $f$ 's which also make physically sensible probabilities, but I won't go into those now. There are other linear combinations that you can ask questions about, but you don't seem to be able to ask questions about an individual  $f$ .

## 6. NEGATIVE PROBABILITIES

Now, for many interacting spins on a lattice we can give a "probability" (the quotes remind us that there is still a question about whether it's a probability) for correlated possibilities:

$$F(s_1, s_2, \dots, s_N) \quad (s_i \in \{++, +-,-+, --\})$$

Next, if I look for the quantum mechanical equation which tells me what the changes of  $F$  are with time, they are exactly of the form that I wrote above for the classical theory:

$$F_{t+1}(\{s\}) = \sum_{\{s'\}} \left[ \prod_i M(s_i|s'_j, s'_{k\dots}) \right] F_t(\{s'\})$$

but now we have  $F$  instead of  $P$ . The  $M(s_i|s'_j, s'_{k\dots})$  would appear to be interpreted as the “probability” per unit time, or per time jump, that the state at  $i$  turns into  $s_i$  when the neighbors are in configuration  $s'$ . If you can invent a probability  $M$  like that, you write the equations for it according to normal logic, those are the correct equations, the real, correct, quantum mechanical equations for this  $F$ , and therefore you'd say, Okay, so I can imitate it with a probabilistic computer!

There's only one thing wrong. These equations unfortunately cannot be so interpreted on the basis of the so-called “probability”, or this probabilistic computer can't simulate them, because the  $F$  is not necessarily positive. Sometimes it's negative! The  $M$ , the “probability” (so-called) of moving from one condition to another is itself not positive; if I had gone all the way back to the  $f$  for a single object, it again is not necessarily positive.

An example of possibilities here are

$$f_{++} = 0.6 \quad f_{+-} = -0.1 \quad f_{-+} = 0.3 \quad f_{--} = 0.2$$

The sum  $f_{++} + f_{+-}$  is 0.5, that's 50% chance of finding the first index positive. The probability of finding the first index negative is the sum  $f_{-+} + f_{--}$  which is also 50%. The probability of finding the second index positive is the sum  $f_{++} + f_{-+}$  which is nine tenths, the probability of finding it negative is  $f_{+-} + f_{--}$  which is one-tenth, perfectly alright, it's either plus or minus. The probability that they match is eight-tenths, the probability that they mismatch is plus two-tenths; every physical probability comes out positive. But the original  $f$ 's are not positive, and therein lies the great difficulty. The only difference between a probabilistic classical world and the equations of the quantum world is that somehow or other it appears as if the probabilities would have to go negative, and that we do not know, as far as I know, how to simulate. Okay, that's the fundamental problem. I don't know the answer to it, but I wanted to explain that if I try my best to make the equations look as near as possible to what would be imitable by a classical probabilistic computer, I get into trouble.

## 7. POLARIZATION OF PHOTONS—TWO-STATES SYSTEMS

I would like to show you why such minus signs cannot be avoided, or at least that you have some sort of difficulty. You probably have all heard this example of the Einstein-Podolsky-Rosen paradox, but I will explain this little example of a physical experiment which can be done, and which has been done, which does give the answers quantum theory predicts, and the answers are really right, there's no mistake, if you do the experiment, it actually comes out. And I'm going to use the example of polarizations of photons, which is an example of a two-state system. When a photon comes, you can say it's either  $x$  polarized or  $y$  polarized. You can find that out by putting in a piece of calcite, and the photon goes through the calcite either out in one direction, or out in another—actually slightly separated, and then you put in some mirrors, that's not important. You get two beams, two places out, where the photon can go. (See Figure 2.)

If you put a polarized photon in, then it will go to one beam called the ordinary ray, or another, the extraordinary one. If you put detectors there you find that each photon that you put in, it either comes out in one or the other 100% of the time, and not half and half. You either find a photon in one or the other. The probability of finding it in the ordinary ray plus the probability of finding it in the extraordinary ray is always 1—you have to have that rule. That works. And further, it's never found at both detectors. (If you might have put two photons in, you could get that, but you cut the intensity down—it's a technical thing, you don't find them in both detectors.)

Now the next experiment: Separation into 4 polarized beams (see Figure 3). You put two calcites in a row so that their axes have a relative angle  $\phi$ , I happen to have drawn the second calcite in two positions, but it doesn't make a difference if you use the same piece or not, as you care. Take the ordinary ray from one and put it through another piece of calcite and look at its ordinary ray, which I'll call the ordinary-ordinary ( $O-O$ ) ray, or look at its extraordinary ray, I have the ordinary-extraordinary ( $O-E$ ) ray. And then the extraordinary ray from the first one comes out as the  $E-O$  ray, and then there's an  $E-E$  ray, alright. Now you can ask what happens.

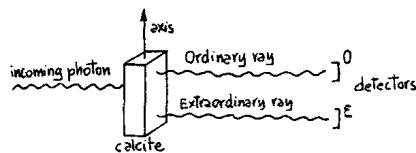


Fig. 2.

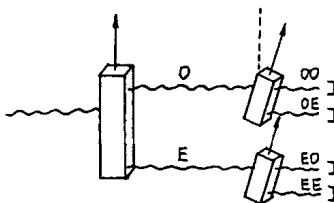


Fig. 3.

You'll find the following. *When a photon comes in, you always find that only one of the four counters goes off.*

If the photon is *O* from the first calcite, then the second calcite gives *O-O* with probability  $\cos^2 \phi$  or *O-E* with the complementary probability  $1 - \cos^2 \phi = \sin^2 \phi$ . Likewise an *E* photon gives a *E-O* with the probability  $\sin^2 \phi$  or an *E-E* with the probability  $\cos^2 \phi$ .

## 8. TWO-PHOTON CORRELATION EXPERIMENT

Let us turn now to the two photon correlation experiment (see Figure 4).

What can happen is that an atom emits two photons in opposite direction (e.g., the  $3s \rightarrow 2p \rightarrow 1s$  transition in the H atom). They are observed simultaneously (say, by you and by me) through two calcites set at  $\phi_1$  and  $\phi_2$  to the vertical. Quantum theory and experiment agree that the probability  $P_{OO}$  that both of us detect an ordinary photon is

$$P_{OO} = \frac{1}{2} \cos^2(\phi_2 - \phi_1)$$

The probability  $P_{EE}$  that we both observe an extraordinary ray is the same

$$P_{EE} = \frac{1}{2} \cos^2(\phi_2 - \phi_1)$$

The probability  $P_{OE}$  that I find *O* and you find *E* is

$$P_{OE} = \frac{1}{2} \sin^2(\phi_2 - \phi_1)$$

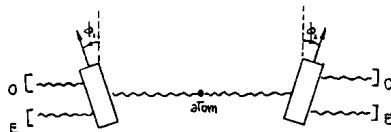


Fig. 4.

and finally the probability  $P_{EO}$  that I measure  $E$  and you measure  $O$  is

$$P_{EO} = \frac{1}{2} \sin^2(\phi_2 - \phi_1)$$

Notice that you can always predict, from your own measurement, what I shall get,  $O$  or  $E$ . For any axis  $\phi_1$  that I chose, just set your axis  $\phi_2$  to  $\phi_1$ , then

$$P_{OE} = P_{EO} = 0$$

and I must get whatever you get.

Let us see now how it would have to be for a *local* probabilistic computer. Photon 1 must be in some condition  $\alpha$  with the probability  $f_\alpha(\phi_1)$ , that determines it to go through as an ordinary ray [the probability it would pass as  $E$  is  $1 - f_\alpha(\phi_1)$ ]. Likewise photon 2 will be in a condition  $\beta$  with probability  $g_\beta(\phi_2)$ . If  $p_{\alpha\beta}$  is the conjoint probability to find the condition pair  $\alpha, \beta$ , the probability  $P_{OO}$  that both of us observe  $O$  rays is

$$P_{OO}(\phi_1, \phi_2) = \sum_{\alpha\beta} p_{\alpha\beta} f_\alpha(\phi_1) g_\beta(\phi_2) \quad \sum_{\alpha\beta} p_{\alpha\beta} = 1$$

likewise

$$P_{OE}(\phi_1, \phi_2) = \sum_{\alpha\beta} p_{\alpha\beta} (1 - f_\alpha(\phi_1)) g_\beta(\phi_2) \quad \text{etc.}$$

The conditions  $\alpha$  determine how the photons go. There's some kind of correlation of the conditions. Such a formula cannot reproduce the quantum results above for any  $p_{\alpha\beta}, f_\alpha(\phi_1), g_\beta(\phi_2)$  if they are real probabilities—that is all positive, although it is easy if they are “probabilities”—negative for some conditions or angles. We now analyze why that is so.

I don't know what kinds of conditions they are, but for any condition the probability  $f_\alpha(\phi)$  of its being extraordinary or ordinary in any direction must be either one or zero. Otherwise you couldn't predict it on the other side. You would be unable to predict with certainty what I was going to get, unless, every time the photon comes here, which way it's going to go is absolutely determined. Therefore, whatever condition the photon is in, there is some hidden inside variable that's going to determine whether it's going to be ordinary or extraordinary. This determination is done deterministically, not probabilistically; otherwise we can't explain the fact that you could predict what I was going to get *exactly*. So let us suppose that something like this happens. Suppose we discuss results just for angles which are multiples of  $30^\circ$ .

On each diagram (Figure 5) are the angles  $0^\circ$ ,  $30^\circ$ ,  $60^\circ$ ,  $90^\circ$ ,  $120^\circ$ , and  $150^\circ$ . A particle comes out to me, and it's in some sort of state, so what it's going to give for  $0^\circ$ , for  $30^\circ$ , etc. are all predicted—determined—by the state. Let us say that in a particular state that is set up the prediction for  $0^\circ$  is that it'll be extraordinary (black dot), for  $30^\circ$  it's also extraordinary, for  $60^\circ$  it's ordinary (white dot), and so on (Figure 5a). By the way, the outcomes are complements of each other at right angles, because, remember, it's always either extraordinary or ordinary; so if you turn  $90^\circ$ , what used to be an ordinary ray becomes the extraordinary ray. Therefore, whatever condition it's in, it has some predictive pattern in which you either have a prediction of ordinary or of extraordinary—three and three—because at right angles they're not the same color. Likewise the particle that comes to you when they're separated must have the same pattern because you can determine what I'm going to get by measuring yours. Whatever circumstances come out, the patterns must be the same. So, if I want to know, Am I going to get white at  $60^\circ$ ? You just measure at  $60^\circ$ , and you'll find white, and therefore you'll predict white, or ordinary, for me. Now each time we do the experiment the pattern may not be the same. Every time we make a pair of photons, repeating this experiment again and again, it doesn't have to be the same as Figure 5a. Let's assume that the next time the experiment my photon will be  $O$  or  $E$  for each angle as in Figure 5c. Then your pattern looks like Figure 5d. But whatever it is, your pattern has to be my pattern exactly—otherwise you couldn't predict what I was going to get exactly by measuring the corresponding angle. And so on. Each time we do the experiment, we get different patterns; and it's easy: there are just six dots and three of them are white, and you chase them around different way—everything can happen. If we measure at the same angle, we always find that with this kind of arrangement we would get the same result.

Now suppose we measure at  $\phi_2 - \phi_1 = 30^\circ$ , and ask, With what probability do we get the same result? Let's first try this example here (Figure 5a, 5b). With what probability would we get the same result, that they're

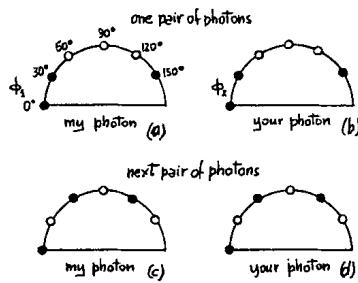


Fig. 5.

both white, or they're both black? The thing comes out like this: suppose I say, After they come out, I'm going to choose a direction at random, I tell you to measure  $30^\circ$  to the right of that direction. Then whatever I get, you would get something different if the neighbors were different. (We would get the same if the neighbors were the same.) What is the chance that you get the same result as me? The chance is the number of times that the neighbor is the same color. If you'll think a minute, you'll find that two thirds of the time, in the case of Figure 5a, it's the same color. The worst case would be black/white/black/white/black/white, and there the probability of a match would be zero (Figure 5c,d). If you look at all eight possible distinct cases, you'll find that the biggest possible answer is two-thirds. You cannot arrange, in a classical kind of method like this, that the probability of agreement at  $30^\circ$  will be bigger than two-thirds. But the quantum mechanical formula predicts  $\cos^2 30^\circ$  (or  $3/4$ )—and experiments agree with this—and therein lies the difficulty.

That's all. That's the difficulty. That's why quantum mechanics can't seem to be imitable by a local classical computer.

I've entertained myself always by squeezing the difficulty of quantum mechanics into a smaller and smaller place, so as to get more and more worried about this particular item. It seems to be almost ridiculous that you can squeeze it to a numerical question that one thing is bigger than another. But there you are—it is bigger than any logical argument can produce, if you have this kind of logic. Now, we say "this kind of logic;" what other possibilities are there? Perhaps there may be no possibilities, but perhaps there are. Its interesting to try to discuss the possibilities. I mentioned something about the possibility of time—of things being affected not just by the past, but also by the future, and therefore that our probabilities are in some sense "illusory." We only have the information from the past, and we try to predict the next step, but in reality it depends upon the near future which we can't get at, or something like that. A very interesting question is the origin of the probabilities in quantum mechanics. Another way of putting things is this: we have an illusion that we can do any experiment that we want. We all, however, come from the same universe, have evolved with it, and don't really have any "real" freedom. For we obey certain laws and have come from a certain past. Is it somehow that we are correlated to the experiments that we do, so that the apparent probabilities don't look like they ought to look if you assume that they are random. There are all kinds of questions like this, and what I'm trying to do is to get you people who think about computer-simulation possibilities to pay a great deal of attention to this, to digest as well as possible the real answers of quantum mechanics, and see if you can't invent a different point of view than the physicists have had to invent to describe this. In fact the physicists have no

good point of view. Somebody mumbled something about a many-world picture, and that many-world picture says that the wave function  $\psi$  is what's real, and damn the torpedos if there are so many variables,  $N^R$ . All these different worlds and every arrangement of configurations are all there just like our arrangement of configurations, we just happen to be sitting in this one. It's possible, but I'm not very happy with it.

So, I would like to see if there's some other way out, and I want to emphasize, or bring the question here, because the discovery of computers and the thinking about computers has turned out to be extremely useful in many branches of human reasoning. For instance, we never really understood how lousy our understanding of languages was, the theory of grammar and all that stuff, until we tried to make a computer which would be able to understand language. We tried to learn a great deal about psychology by trying to understand how computers work. There are interesting philosophical questions about reasoning, and relationship, observation, and measurement and so on, which computers have stimulated us to think about anew, with new types of thinking. And all I was doing was hoping that the computer-type of thinking would give us some new ideas, if any are really needed. I don't know, maybe physics is absolutely OK the way it is. The program that Fredkin is always pushing, about trying to find a computer simulation of physics, seem to me to be an excellent program to follow out. He and I have had wonderful, intense, and interminable arguments, and my argument is always that the real use of it would be with quantum mechanics, and therefore full attention and acceptance of the quantum mechanical phenomena—the challenge of explaining quantum mechanical phenomena—has to be put into the argument, and therefore these phenomena have to be understood very well in analyzing the situation. And I'm not happy with all the analyses that go with just the classical theory, because nature isn't classical, dammit, and if you want to make a simulation of nature, you'd better make it quantum mechanical, and by golly it's a wonderful problem, because it doesn't look so easy. Thank you.

## 9. DISCUSSION

*Question:* Just to interpret, you spoke first of the probability of A given B, versus the probability of A and B jointly—that's the probability of one observer seeing the result, assigning a probability to the other; and then you brought up the paradox of the quantum mechanical result being  $3/4$ , and this being  $2/3$ . Are those really the same probabilities? Isn't one a joint probability, and the other a conditional one?

*Answer:* No, they are the same.  $P_{OO}$  is the *joint probability* that both you and I observe an ordinary ray, and  $P_{EE}$  is the *joint probability* for two

extraordinary rays. The probability that our observations match is

$$P_{OO} + P_{EE} = \cos^2 30^\circ = 3/4$$

*Question:* Does it in some sense depend upon an assumption as to how much information is accessible from the photon, or from the particle? And second, to take your question of prediction, your comment about predicting, is in some sense reminiscent of the philosophical question, Is there any meaning to the question of whether there is free will or predestination? namely, the correlation between the observer and the experiment, and the question there is, Is it possible to construct a test in which the prediction could be reported to the observer, or instead, has the ability to represent information already been used up? And I suspect that you may have already used up all the information so that prediction lies outside the range of the theory.

*Answer:* All these things I don't understand; deep questions, profound questions. However physicists have a kind of a dopy way of avoiding all of these things. They simply say, now look, friend, you take a pair of counters and you put them on the side of your calcite and you count how many times you get this stuff, and it comes out 75% of the time. Then you go and you say, Now can I imitate that with a device which is going to produce the same results, and which will operate locally, and you try to invent some kind of way of doing that, and if you do it in the ordinary way of thinking, you find that you can't get there with the same probability. Therefore some new kind of thinking is necessary, but physicists, being kind of dull minded, only look at nature, and don't know how to think in these new ways.

*Question:* At the beginning of your talk, you talked about discretizing various things in order to go about doing a real computation of physics. And yet it seems to me that there are some differences between things like space and time, and probability that might exist at some place, or energy, or some field value. Do you see any reason to distinguish between quantization or discretizing of space and time, versus discretizing any of the specific parameters or values that might exist?

*Answer:* I would like to make a few comments. You said quantizing or discretizing. That's very dangerous. Quantum theory and quantizing is a very specific type of theory. Discretizing is the right word. Quantizing is a different kind of mathematics. If we talk about discretizing... of course I pointed out that we're going to have to change the laws of physics. Because the laws of physics as written now have, in the classical limit, a continuous variable everywhere, space and time. If, for example, in your theory you were going to have an electric field, then the electric field could not have (if it's going to be imitable, computable by a finite number of elements) an

infinite number of possible values, it'd have to be digitized. You might be able to get away with a theory by redescribing things without an electric field, but supposing for a moment that you've discovered that you can't do that and you want to describe it with an electric field, then you would have to say that, for example, when fields are smaller than a certain amount, they aren't there at all, or something. And those are very interesting problems, but unfortunately they're not good problems for classical physics because if you take the example of a star a hundred light years away, and it makes a wave which comes to us, and it gets weaker, and weaker, and weaker, and weaker, the electric field's going down, down, down, how low can we measure? You put a counter out there and you find "clunk," and nothing happens for a while, "clunk," and nothing happens for a while. It's not discretized at all, you never can measure such a tiny field, you don't find a tiny field, you don't have to imitate such a tiny field, because the world that you're trying to imitate, the physical world, is not the classical world, and it behaves differently. So the particular example of discretizing the electric field, is a problem which I would not see, as a physicist, as fundamentally difficult, because it will just mean that your field has gotten so small that I had better be using quantum mechanics anyway, and so you've got the wrong equations, and so you did the wrong problem! That's how I would answer that. Because you see, if you would imagine that the electric field is coming out of some 'ones' or something, the lowest you could get would be a full one, but that's what we see, you get a full photon. All these things suggest that it's really true, somehow, that the physical world is representable in a discretized way, because every time you get into a bind like this, you discover that the experiment does just what's necessary to escape the trouble that would come if the electric field went to zero, or you'd never be able to see a star beyond a certain distance, because the field would have gotten below the number of digits that your world can carry.

**A NEW INFLATIONARY UNIVERSE SCENARIO: A POSSIBLE SOLUTION  
OF THE HORIZON, FLATNESS, HOMOGENEITY, ISOTROPY AND  
PRIMORDIAL MONPOLE PROBLEMS**

A.D. LINDE

*Lebedev Physical Institute, Moscow 117924, USSR*

Received 29 October 1981

A new inflationary universe scenario is suggested, which is free of the shortcomings of the previous one and provides a possible solution of the horizon, flatness, homogeneity and isotropy problems in cosmology, and also a solution of the primordial monopole problem in grand unified theories.

There is now considerable interest in the cosmological consequences of symmetry breaking phase transitions, which occur in grand unified theories (GUTs) with the decrease of temperature at the very early stages of the evolution of the universe [1–3]. These phase transitions typically are strongly first order [4,5]. The lifetime of the supercooled symmetric phase  $\varphi = 0$  ( $\varphi$  is the Higgs scalar field which breaks the symmetry) in some theories may be extremely large [2,3,6–8]. In that case the energy-momentum tensor of particles  $\sim T^4$  in the phase  $\varphi = 0$  almost vanishes in the course of the expansion of the universe, and the total energy-momentum tensor reduces to the vacuum stress tensor (cosmological term)  $T_{\mu\nu}^{\text{vac}} = g_{\mu\nu} V(0)$ , where  $V(\varphi)$  is the effective potential of the theory at vanishing temperature [2,3]. This leads to an exponentially fast expansion of the universe,  $a \sim e^{Ht}$ . Here  $a$  is the scale factor, and  $H$  is the Hubble constant at that time,

$$H = [(8\pi/3M_P^2)V(0)]^{1/2},$$

where  $M_P \approx 10^{19}$  GeV is the Planck mass [9]. Then at some comparatively small temperature  $T_c$  the symmetry breaking phase transition takes place, all the vacuum energy  $V(0)$  transforms into thermal energy [2,3], the universe is reheated up to the temperature  $T_1 \approx V_{(0)}^{1/4}$ , and its further evolution proceeds in a standard way [10] <sup>#1</sup>.

A most detailed discussion of this scenario is con-

tained in a very interesting paper of Guth [12], where it is shown that the existence of a sufficiently long period of exponential expansion (inflation) in the early universe would provide a natural solution of the horizon and flatness problems in cosmology and of the primordial monopole problem in grand unified theories [13].

Unfortunately, however, this scenario in the form suggested in ref. [12] leads to some unacceptable consequences, recognized by Guth himself and by other authors who have studied this problem later, see e.g. refs. [8,14–18]. The phase transition from the symmetric vacuum state  $\varphi = 0$  to the asymmetric state  $\varphi = \varphi_0$  proceeds by creation and subsequent expansion of bubbles containing some nonvanishing fields  $\varphi$ . In ref. [12] it was implicitly assumed that inside these bubbles the scalar field  $\varphi$  rapidly grows to  $\varphi = \varphi_0$ , all energy of the bubbles becomes concentrated in their walls and thermalization occurs only after the collision of the walls. If this qualitative picture were correct, the exponential expansion would be finished at the temperature  $T_c$ , at which the phase transition occurs. For the flatness problem to be solved the universe (the scale factor  $a$ ) should grow at least  $10^{28}$

<sup>#1</sup> For an alternative scenario of the exponential expansion at the very early stages of the evolution of the universe, which may occur due to quantum gravity effects, see ref. [11].

times during the exponential expansion period [12]. Since at this period the value of  $aT$  is constant, the critical temperature  $T_c$  should be  $10^{28}$  times smaller than the temperature  $T_0$ , at which the exponential expansion starts. In the simplest SU(5) model [19]  $T_0 \sim 10^{14}$  GeV, so that

$$T_c \lesssim 10^{-14} \text{ GeV} \sim 0.1 \text{ K}. \quad (1)$$

No GUTs with such a fantastically small value of the critical temperature have been suggested.

There is also another problem with the above mentioned scenario. If the bubble wall collisions are necessary for the reheating of the universe, then after such a phase transition the universe becomes greatly inhomogeneous and anisotropic, which would contradict cosmological data [8,12,17,18].

In the present paper we would like to suggest an improved inflationary universe scenario, which is free of the above mentioned difficulties. With this purpose we shall consider the phase transitions in GUTs with the Coleman–Weinberg mechanism of symmetry breaking [20]. Phase transitions in such theories have been studied recently by many authors [14,15,21–23]. In our opinion, however, several important features of these phase transitions have escaped their attention. A detailed discussion of the phase transitions in GUTs with the Coleman–Weinberg mechanism of symmetry breaking will be contained in a subsequent publication. Here we shall only outline the main idea, which is essential for the understanding of the new inflationary universe scenario.

For definiteness let us consider the SU(5) grand unified theory [19], though most of what will be discussed here will not depend on the details of the model under consideration. The one-loop effective potential for the symmetry breaking  $SU(5) \rightarrow SU(3) \times SU(2) \times U(1)$  in the Coleman–Weinberg version of this model at finite temperature  $T$  is [1–3, 14,15]

$$\begin{aligned} V(\varphi, T) &= (18T^4/\pi^2) \\ &\times \int_0^\infty dx x^2 \ln\{1 - \exp[-(x^2 + 25g^2\varphi^2/8T^2)^{1/2}]\} \\ &+ (5625/512\pi^2)g^4(\varphi^4 \ln(\varphi/\varphi_0) - \varphi^4/4 + \varphi_0^4/4), \end{aligned} \quad (2)$$

where  $\varphi_0 \sim 10^{14}–10^{15}$  GeV,  $g^2 \sim 1/3$  is the gauge

coupling constant. At  $T \gg \varphi_0$  the symmetry in this theory was restored,  $\varphi = 0$  [1–3]. With a decrease of temperature the absolute minimum of  $V(\varphi)$  appears at  $\varphi \approx \varphi_0$ . However at any  $T \neq 0$  the point  $\varphi = 0$  remains a local minimum of  $V(\varphi, T)$ , since near  $\varphi = 0$

$$\begin{aligned} V(\varphi, T) &= \frac{75}{16}g^2T^2\varphi^2 \\ &- (5625/512\pi^2)g^4\varphi^4 \ln(M_x/T) + (9/32\pi^2)M_x^4, \end{aligned} \quad (3)$$

where  $M_x^2 = \frac{25}{8}g^2\varphi_0^2$ . The phase transition with symmetry breaking proceeds from a strongly supercooled state  $\varphi = 0$  at temperature  $T_c$ , which is many orders of magnitude smaller than  $\varphi_0$  [14,15,21–23]. The shape of the potential  $V(\varphi, T)$  for  $T \ll \varphi_0$  is shown in fig. 1. The phase transition begins with the formation of bubbles of the field  $\varphi$ , which is a tunneling process [6]. One may argue that for  $T_c \ll \varphi_0$  this process does not depend on the properties of  $V(\varphi, T_c)$  at  $\varphi \sim \varphi_0$ , and the maximal value of the field  $\varphi$  inside the bubble immediately after its formation should be of the order of  $\varphi_1$ , where

$$V(\varphi_1, T_c) = V(0, T_c), \quad (4)$$

$\varphi_1 \ll \varphi_0$ , see fig. 1. Indeed, a detailed study of the bubble formation in this theory performed in refs. [24,25] by means of computer calculations shows that at the moment of the bubble formation the maximal value of the field  $\varphi$  inside the bubble equals approximately  $3\varphi_1$ . Therefore inside the bubble

$$\varphi \lesssim 3\varphi_1 = \frac{12\pi T_c}{5g} \left( \frac{2}{3 \ln(M_x/T_c)} \right)^{1/2} \ll \varphi_0. \quad (5)$$

This means that the (negative) mass squared of the

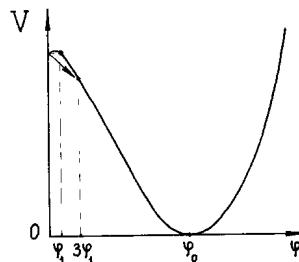


Fig. 1. Effective potential in the Coleman–Weinberg theory for  $T \ll \varphi_0$ . The arrow indicates the direction of the tunneling with bubble formation.

field  $\varphi$  inside the bubble is

$$-m^2 = -\frac{2}{15} d^2 V/d\varphi^2 \lesssim 75g^2 T_c^2 \sim 25 T_c^2. \quad (6)$$

After the bubble formation the field  $\varphi$  inside the bubble gradually grows up to its equilibrium value  $\varphi(T_1) \sim \varphi_0$ . At the first stages of this process the field  $\varphi$  grows approximately as  $e^{mt}$ . Therefore it approaches its equilibrium value  $\varphi(T_1)$  only after some period of time  $\tau \gtrsim m^{-1} \sim 0.2T_c^{-1}$ . A more complete investigation shows that  $\tau$  is several times greater than  $m^{-1}$ ; here for simplicity we shall take as an estimate

$$\tau \sim T_c^{-1}. \quad (7)$$

It can also be easily shown that during most of this period the field  $\varphi$  inside the bubble remains much less than  $\varphi_0$ . Therefore during some time of the order of  $\tau \sim T_c^{-1}$  the vacuum energy density  $V(\varphi)$  remains almost equal to  $V(0)$ , and the part of the universe inside the bubble expands exponentially just as it expanded before the bubble creation. This simple observation has very important consequences for the theory of the phase transitions in the Coleman–Weinberg model.

Let us suppose that the phase transition in the Coleman–Weinberg SU(5) theory occurs at  $T_c \sim 2 \times 10^6$  GeV, as it was claimed in ref. [15]. From eq. (3) it follows that with the parameters of the theory used in ref. [15] ( $M_x \sim 6 \times 10^{14}$  GeV) the Hubble constant

$$H = [(8\pi/3M_p^2)V(0)]^{1/2}$$

is equal to  $1.5 \times 10^{10}$  GeV. Therefore during the exponential expansion period  $\tau \sim T_c^{-1}$  the universe should grow  $e^{H\tau}$  times, where

$$e^{H\tau} \sim e^{H/T_c} \sim e^{7500} \sim 10^{3260}. \quad (8)$$

A typical size of the bubble at the moment of its creation is  $O(T_c^{-1}) \sim 10^{-20}$  cm [25]. After the period of the exponential expansion this bubble will have a size of

$$10^{-20} \cdot e^{H\tau} \text{ cm} \sim 10^{3240} \text{ cm},$$

which is much greater than the size of the observable part of the universe  $l \sim 10^{28}$  cm. Therefore the whole observable part of the universe is contained *inside one bubble*, so we see no inhomogeneities caused by the wall collisions. After some time of the order of  $\tau$

after the bubble creation all the vacuum energy density  $V(0)$  transforms into thermal energy  $\sim T_1^4$ , where in our model  $T_1 \approx 0.15M_x \sim 10^{14}$  GeV. However the thermalization occurs now not due to the wall collisions, but due to the interactions of particles created by the classical homogeneous field  $\varphi$ , convergently oscillating near its equilibrium value  $\varphi(T_1) \approx \varphi_0$  with a frequency of about  $10^{14}$  GeV.

One can easily verify that the size of the particle horizon at the time of the phase transition was much greater than the size of the bubble  $\sim T_c^{-1}$ , i.e. all points inside the bubble were causally connected. After the exponential expansion period this causally connected domain covers the whole observable part of the universe, which solves the horizon problem [12].

Now let us remember that particle creation in the very early universe in general cannot make it completely isotropic, but makes it quasi-isotropic [10], i.e. locally isotropic in small domains of space of the size of the same order as or greater than the Planck length  $l_p \sim 10^{-33}$  cm  $\sim M_p^{-1}$  at the Planck time  $t_p \sim 10^{-43}$  s, when the temperature was  $T_p \sim M_p \sim 10^{19}$  GeV. Since before the phase transition the quantity  $aT$  was constant inside each isotropic domain of the universe, at the moment of the phase transition the typical size of the isotropic domain exceeds the bubble size  $\sim T_c^{-1}$ . Therefore the space-time inside the bubble was isotropic and the exponential expansion extends this isotropy to the whole observable part of the universe. (Moreover, the remaining small anisotropy inside the bubble decreases rapidly during the exponential expansion period [18].) This may solve the long-standing problem of the space-time isotropy in our universe.

Density fluctuations inside the bubble immediately after its formation are negligibly small as compared with  $V(0)$ , i.e. the space inside the bubble is almost homogeneous. Then the exponential expansion extends this homogeneity to the whole observable part of the universe, which explains the large-scale homogeneity of the universe.

One may argue that it is not very good to obtain an absolutely homogeneous universe, since in that case it would be difficult to understand the origin of galaxies. However, as will be explained in a separate publication (see also refs. [7,26]), the necessary inhomogeneities may be generated after a subsequent

phase transition with a smaller degree of supercooling. Moreover, as is shown in ref. [27], the perturbations necessary for galaxy formation arise due to quantum gravity effects just after phase transitions of the type considered above in GUTs with the unification scale  $\Lambda \sim 10^{17} - 10^{18}$  GeV, which is not unrealistic [28].

From our results it follows that the size of the universe  $l_1$  after the phase transition should exceed  $10^{3240}$  cm, and the temperature  $T_1$  is of the order of  $10^{14}$  GeV. Therefore the total entropy of the universe should exceed  $(l_1 T_1)^3 \sim 10^{10000}$ , which explains why the total entropy of the universe exceeds  $10^{85}$  and simultaneously solves the flatness problem [12,21].

It is known that the primordial monopoles in GUTs are created only in the points, in which bubbles with different types of Higgs field  $\varphi$  collide [13]. Therefore in our scenario no monopoles are created in the observable part of the universe, which solves the primordial monopole problem in GUTs [13]. For the same reason there will be no domain walls in the observable part of the universe in the theories with broken discrete symmetries [29], and in particular in the theories with spontaneously broken  $CP$  invariance. This helps to solve the problem of the baryon asymmetry of the universe [30]. Moreover, in the scenario under consideration there appears an additional source of baryon asymmetry. The standard mechanism is connected with the decay of the X, Y bosons and Higgs mesons, which appear in the course of the reheating of the universe during the phase transition. It is clear, however, that the baryon asymmetry generated by the decay of the Higgs mesons may be generated by the decay of the classical Higgs field vacuum  $\varphi = 0$  as well. Note also that, whereas in a standard scenario the particles created by the decay of the X, Y and Higgs mesons are only a few percent of the total amount of particles [30], in our case all particles which appear after the phase transition are created by the oscillating classical Higgs field  $\varphi$  during the phase transition.

The new inflationary universe scenario discussed above is, of course, oversimplified. To get a complete scenario, one should analyse the phase transitions in the Coleman–Weinberg model more accurately taking into account the renormalization group equation [21] and the nonperturbative effects [15]. This anal-

ysis should be performed simultaneously with the investigation of the effects connected with the nonvanishing curvature and rapid expansion of the universe, which become important for  $H > T_c$ . One may ask e.g. whether it is possible for the universe to be in a state with temperature  $T_c$  smaller than the Hawking temperature  $T_H = H/2\pi$ , to which consequences the terms  $\sim R\varphi^2$  in the effective potential may lead etc. [23]. An investigation of these problems is rather involved and will be contained in a separate publication. Our preliminary result is that there exists an *improved Coleman–Weinberg theory*, in which  $d^2V/d\varphi^2 = 0$  at  $\varphi = 0$  not in Minkowski space, but rather in de Sitter space with the curvature  $R$  determined by the vacuum energy density  $V(\varphi)$  at the symmetric point  $\varphi = 0$ . In some versions of this theory the phase transition with symmetry breaking occurs due to non-perturbative effects [15]. The kinetics of this phase transition is somewhat more complicated than that described above, but the main feature of the new inflationary universe scenario remains intact: The field  $\varphi$  approaches its equilibrium value  $\varphi(T_1) \sim \varphi_0$  during the period  $\tau \gg H^{-1}$ , which just leads to the desirable inflation of the universe discussed in the present paper

I would like to express my deep gratitude to G.V. Chibisov, P.C.W. Davies, V.P. Frolov, L.P. Grishchuk, S.W. Hawking, R.E. Kallosh, D.A. Kirzhnits, V.F. Mukhanov, V.A. Rubakov, A.A. Starobinsky, A.V. Veryaskin and Ya.B. Zeldovich for many enlightening discussions.

### References

- [1] D.A. Kirzhnits, JETP Lett. 15 (1972) 529; D.A. Kirzhnits and A.D. Linde, Phys. Lett. 42B (1972) 471; S. Weinberg, Phys. Rev. D9 (1974) 3357; L. Dolan and R. Jackiw, Phys. Rev. D9 (1974) 3320; D.A. Kirzhnits and A.D. Linde, Zh. Eksp. Teor. Fiz. 67 (1974) 1263 [Sov. Phys. JETP 40 (1975) 628].
- [2] D.A. Kirzhnits and A.D. Linde, Ann. Phys. (NY) 101 (1976) 195.
- [3] A.D. Linde, Rep. Prog. Phys. 42 (1979) 389.
- [4] A.D. Linde, in: Statistical mechanics of quarks and hadrons, ed. H. Satz (North-Holland, Amsterdam, 1981) p. 385; Phys. Lett. 99B (1981) 391.
- [5] M. Daniel, Phys. Lett. 98B (1981) 371.
- [6] A.D. Linde, Phys. Lett. 70B (1977) 306; 100B (1981) 37.

- [7] K. Sato, Mon. Not. R. Astron. Soc. 195 (1981) 467.
- [8] A.H. Guth and E. Weinberg, Phys. Rev. D23 (1981) 876.
- [9] R. Tolman, Relativity, thermodynamics and cosmology (Clarendon, Oxford, 1969).
- [10] Ya.B. Zeldovich and I.D. Novikov, Structure and evolution of the universe (Nauka, Moscow, 1975).
- [11] A.A. Starobinsky, Phys. Lett. 91B (1980) 100;  
Ya.B. Zeldovich, Pis'ma Astron. Zh. 7 (1981) 579.
- [12] A.H. Guth, Phys. Rev. D23 (1981) 347.
- [13] Ya.B. Zeldovich and M.Yu. Khlopov, Phys. Lett. 79B (1978) 239;  
J.P. Preskill, Phys. Rev. Lett. 43 (1979) 1365.
- [14] G.P. Cook and K.T. Mahanthappa, Phys. Rev. D23 (1981) 1321.
- [15] A. Billoire and K. Tamvakis, CERN preprint TH. 3019 (1981);  
K. Tamvakis and C.E. Vayonakis, CERN preprints TH. 3108, TH. 3128 (1981).
- [16] Q. Shafi, CERN preprint TH. 3143 (1981).
- [17] S.W. Hawking, I.G. Moss and J.M. Stewart, DAMTP preprint (1981).
- [18] J.D. Barrow and M.S. Turner, Nature 292 (1981) 35.
- [19] H. Georgi and S.L. Glashow, Phys. Rev. Lett. 32 (1974) 389.
- [20] S. Coleman and E. Weinberg, Phys. Rev. D7 (1973) 1888.
- [21] V.G. Lapchinsky, V.A. Rubakov and A.V. Veryaskin, IYAI preprint (1981).
- [22] M. Sher, Univ. of California preprint NSF-ITP-81-15 (1981).
- [23] L.F. Abbott, Nucl. Phys. B185 (1981) 233;  
P. Hut and F.R. Klinkhamer, Phys. Lett. 104B (1981) 439.
- [24] E. Brezin and G. Parisi, J. Stat. Phys. 19 (1978) 269.
- [25] A.D. Linde, Decay of the false vacuum at finite temperature, Lebedev Phys. Inst. preprint (1981).
- [26] Ya.B. Zeldovich, Mon. Not. R. Astron. Soc. 192 (1980) 663;  
A. Vilenkin, Phys. Rev. Lett. 46 (1981) 1169; Tufts Univ. preprint (1981).
- [27] G.V. Chibisov and V.P. Mukhanov, Lebedev Phys. Inst. preprint No. 198 (1981); Mon. Not. R. Astron. Soc., to be published;  
D.A. Kompaneets, V.N. Lukash and I.D. Novikov, Space Research Inst. preprint No. 652 (1981).
- [28] S. Dimopoulos, S. Raby and F. Wilczek, Stanford Univ. preprint NSF-ITP-81-31 (1981);  
S. Dimopoulos and H. Georgi, Harvard Univ. preprint HUTP-81/A022 (1981).
- [29] Ya.B. Zeldovich, I.Yu. Kobzarev and L.B. Okun, Sov. Phys. JETP 40 (1975) 1.
- [30] A.D. Sakharov, Pis'ma Zh. Eksp. Teor. Fiz. 5 (1967) 32;  
M. Yoshimura, Phys. Rev. Lett. 41 (1978) 281;  
S. Dimopoulos and L. Susskind, Phys. Rev. D18 (1978) 4500;  
J. Ellis, M.K. Gaillard and D.V. Nanopoulos, Phys. Lett. 80B (1979) 360;  
S. Weinberg, Phys. Rev. Lett. 42 (1979) 859.



# Quantum theory, the Church-Turing principle and the universal quantum computer

DAVID DEUTSCH\*

Appeared in *Proceedings of the Royal Society of London A* **400**, pp. 97-117 (1985)<sup>†</sup>

(Communicated by R. Penrose, F.R.S. — Received 13 July 1984)

## Abstract

It is argued that underlying the Church-Turing hypothesis there is an implicit physical assertion. Here, this assertion is presented explicitly as a physical principle: ‘every finitely realizable physical system can be perfectly simulated by a universal model computing machine operating by finite means’. Classical physics and the universal Turing machine, because the former is continuous and the latter discrete, do not obey the principle, at least in the strong form above. A class of model computing machines that is the quantum generalization of the class of Turing machines is described, and it is shown that quantum theory and the ‘universal quantum computer’ are compatible with the principle. Computing machines resembling the universal quantum computer could, in principle, be built and would have many remarkable properties not reproducible by any Turing machine. These do not include the computation of non-recursive functions, but they do include ‘quantum parallelism’, a method by which certain probabilistic tasks can be performed faster by a universal quantum computer than by any classical restriction of it. The intuitive explanation of these properties places an intolerable strain on all interpretations of quantum theory other than Everett’s. Some of the numerous connections between the quantum theory of computation and the rest of physics are explored. Quantum complexity theory allows a physically more reasonable definition of the ‘complexity’ or ‘knowledge’ in a physical system than does classical complexity theory.

---

\*Current address: Centre for Quantum Computation, Clarendon Laboratory, Department of Physics, Parks Road, OX1 3PU Oxford, United Kingdom. Email: david.deutsch@qubit.org

<sup>†</sup>This version (Summer 1999) was edited and converted to L<sup>A</sup>T<sub>E</sub>X by Wim van Dam at the Centre for Quantum Computation. Email: wimvdam@qubit.org

# 1 Computing machines and the Church-Turing principle

The theory of computing machines has been extensively developed during the last few decades. Intuitively, a computing machine is any physical system whose dynamical evolution takes it from one of a set of ‘input’ states to one of a set of ‘output’ states. The states are labelled in some canonical way, the machine is prepared in a state with a given input label and then, following some motion, the output state is measured. For a classical deterministic system the measured output label is a definite function  $f$  of the prepared input label; moreover the value of that label can in principle be measured by an outside observer (the ‘*user*’) and the machine is said to ‘compute’ the function  $f$ .

Two classical deterministic computing machines are ‘computationally equivalent’ under given labellings of their input and output states if they compute the same function under those labellings. But quantum computing machines, and indeed classical stochastic computing machines, do not ‘compute functions’ in the above sense: the output state of a stochastic machine is random with only the probability distribution function for the possible outputs depending on the input state. The output state of a quantum machine, although fully determined by the input state is not an observable and so the user cannot in general discover its label. Nevertheless, the notion of computational equivalence can be generalized to apply to such machines also.

Again we define computational equivalence *under given labellings*, but it is now necessary to specify more precisely what is to be labelled. As far as the input is concerned, labels must be given for each of the possible ways of preparing the machine, which correspond, by definition, to all the possible input states. This is identical with the classical deterministic case. However, there is an asymmetry between input and output because there is an asymmetry between preparation and measurement: whereas a quantum system can be prepared in any desired permitted input state, measurement cannot in general determine its output state; instead one must measure the value of some observable. (Throughout this paper I shall be using the Schrödinger picture, in which the quantum state is a function of time but observables are constant operators.) Thus what must be labelled is the set of ordered pairs consisting of an output observable and a possible measured value of that observable (in quantum theory, a Hermitian operator and one of its eigenvalues). Such an ordered pair contains, in effect, the specification of a possible experiment that could be made on the output, together with a possible result of that experiment.

Two computing machines are computationally equivalent under given labellings if in any possible experiment or sequence of experiments in which their inputs were prepared equivalently under the input labellings, and observables corresponding to each other under the output labellings were measured, the measured values of these observables for the two machines would be statistically indistinguishable. That is, the probability distribution functions for the outputs of the two machines would be identical.

In the sense just described, a given computing machine  $\mathcal{M}$  computes at most one function. However, there ought to be no fundamental difference between altering the input state in which  $\mathcal{M}$  is prepared, and altering systematically the constitution of  $\mathcal{M}$  so that it becomes a different machine  $\mathcal{M}'$  computing a different function. To formalize such operations, it is often useful to consider machines with two inputs, the preparation of one constituting a ‘program’ determining which function of the other is to be computed. To each such machine  $\mathcal{M}$  there corresponds a set  $C(\mathcal{M})$  of ‘ $\mathcal{M}$ -computable functions’. A function  $f$  is  $\mathcal{M}$ -computable if  $\mathcal{M}$  can compute  $f$  when prepared with some program.

The set  $C(\mathcal{M})$  can be enlarged by enlarging the set of changes in the constitution of  $\mathcal{M}$  that are labelled as possible  $\mathcal{M}$ -programs. Given two machines  $\mathcal{M}$  and  $\mathcal{M}'$  it is possible to construct a composite machine whose set of computable functions contains the union of  $C(\mathcal{M})$  and  $C(\mathcal{M}')$ .

There is no purely logical reason why one could not go on *ad infinitum* building more powerful

computing machines, nor why there should exist any function that is outside the computable set of every physically possible machine. Yet although logic does not forbid the physical computation of arbitrary functions, it seems that physics does. As is well known, when designing computing machines one rapidly reaches a point when adding additional hardware does not alter the machine's set of computable functions (under the idealization that the memory capacity is in effect unlimited); moreover, for functions from the integers  $\mathbb{Z}$  to themselves the set  $C(\mathcal{M})$  is always contained in  $C(\mathcal{T})$ , where  $\mathcal{T}$  is Turing's universal computing machine (Turing 1936).  $C(\mathcal{T})$  itself, also known as the set of recursive functions, is denumerable and therefore infinitely smaller than the set of all functions from  $\mathbb{Z}$  to  $\mathbb{Z}$ .

Church (1936) and Turing (1936) conjectured that these limitations on what can be computed are not imposed by the state-of-the-art in designing computing machines, nor by our ingenuity in constructing models for computation, but are universal. This is called the 'Church-Turing hypothesis'; according to Turing,

$$\begin{aligned} \text{Every 'function which would naturally be regarded as computable' can be} \\ \text{computed by the universal Turing machine.} \end{aligned} \tag{1.1}$$

The conventional, non-physical view of (1.1) interprets it as the quasi-mathematical conjecture that all possible formalizations of the intuitive mathematical notion of 'algorithm' or 'computation' are equivalent to each other. But we shall see that it can also be regarded as asserting a new physical principle, which I shall call the Church-Turing principle to distinguish it from other implications and connotations of the conjecture (1.1).

Hypothesis (1.1) and other formulations that exist in the literature (see Hofstadter (1979) for an interesting discussion of various versions) are very vague by comparison with physical principles such as the laws of thermodynamics or the gravitational equivalence principle. But it will be seen below that my statement of the Church-Turing principle (1.2) is manifestly physical, and unambiguous. I shall show that it has the same epistemological status as other physical principles.

I propose to reinterpret Turing's 'functions which would naturally be regarded as computable' as the functions which may in principle be computed by a real physical system. For it would surely be hard to regard a function 'naturally' as computable if it could not be computed in Nature, and conversely. To this end I shall define the notion of '*perfect simulation*'. A computing machine  $\mathcal{M}$  is capable of perfectly simulating a physical system  $\mathcal{S}$ , under a given labelling of their inputs and outputs, if there exists a program  $\pi(\mathcal{S})$  for  $\mathcal{M}$  that renders  $\mathcal{M}$  computationally equivalent to  $\mathcal{S}$  under that labelling. In other words,  $\pi(\mathcal{S})$  converts  $\mathcal{M}$  into a 'black box' functionally indistinguishable from  $\mathcal{S}$ .

I can now state the physical version of the Church-Turing principle:

$$\begin{aligned} \text{'Every finitely realizable physical system can be perfectly simulated by a} \\ \text{universal model computing machine operating by finite means'.} \end{aligned} \tag{1.2}$$

This formulation is both better defined and more physical than Turing's own way of expressing it (1.1), because it refers exclusively to objective concepts such as 'measurement', 'preparation' and 'physical system', which are already present in measurement theory. It avoids terminology like 'would naturally be regarded', which does not fit well into the existing structure of physics.

The 'finitely realizable physical systems' referred to in (1.2) must include any physical object upon which experimentation is possible. The 'universal computing machine' on the other hand, need only be an idealized (but theoretically permitted) finitely specifiable model. The labellings implicitly referred to in (1.2) must also be finitely specifiable.

The reference in (1.1) to a specific universal computing machine (Turing's) has of necessity been replaced in (1.2) by the more general requirement that this machine operate 'by finite means'. 'Finite

means' can be defined axiomatically, without restrictive assumptions about the form of physical laws (cf. Gandy 1980). If we think of a computing machine as proceeding in a sequence of steps whose duration has a non-zero lower bound, then it operates by 'finite means' if (i) only a finite subsystem (though not always the same one) is in motion during anyone step, and (ii) the motion depends only on the state of a finite subsystem, and (iii) the rule that specifies the motion can be given finitely in the mathematical sense (for example as an integer). Turing machines satisfy these conditions, and so does the universal quantum computer  $\mathcal{Q}$  (see §2).

The statement of the Church-Turing principle (1.2) is stronger than what is strictly necessitated by (1.1). Indeed it is so strong that it is *not* satisfied by Turing's machine in classical physics. Owing to the continuity of classical dynamics, the possible states of a classical system necessarily form a continuum. Yet there are only countably many ways of preparing a finite input for  $\mathcal{T}$ . Consequently  $\mathcal{T}$  cannot perfectly simulate any classical dynamical system. (The well studied theory of the 'simulation' of continuous systems by  $\mathcal{T}$  concerns itself not with perfect simulation in my sense but with successive discrete approximation.) In §3, I shall show that it is consistent with our present knowledge of the interactions present in Nature that every real (dissipative) finite physical system can be perfectly simulated by the universal quantum computer  $\mathcal{Q}$ . Thus quantum theory is compatible with the strong form (1.2) of the Church-Turing principle.

I now return to my argument that (1.2) is an empirical assertion. The usual criterion for the empirical status of a theory is that it be experimentally falsifiable (Popper 1959), i.e. that there exist potential observations that would contradict it. However, since the deeper theories we call 'principles' make reference to experiment only *via* other theories, the criterion of falsifiability must be applied indirectly in their case. The principle of conservation of energy, for example, is not in itself contradicted by any conceivable observation because it contains no specification of how to measure energy. The third law of thermodynamics whose form

$$\begin{aligned} & \text{'No finite process can reduce the entropy or temperature of a finitely realizable} \\ & \text{physical system to zero'} \end{aligned} \tag{1.3}$$

bears a certain resemblance to that of the Church-Turing principle, is likewise not directly refutable: no temperature measurement of finite accuracy could distinguish absolute zero from an arbitrarily small positive temperature. Similarly, since the number of possible programs for a universal computer is infinite, no experiment could in general verify that none of them can simulate a system that is thought to be a counter-example to (1.2).

But all this does not place 'principles' outside the realm of empirical science. On the contrary, they are essential frameworks within which directly testable theories are formulated. Whether or not a given physical theory contradicts a principle is first determined by logic alone. Then, if the directly testable theory survives crucial tests but contradicts the principle, that principle is deemed to be refuted, albeit indirectly. If all known experimentally corroborated theories satisfy a restrictive principle, then that principle is corroborated and becomes, on the one hand, a guide in the construction of new theories, and on the other, a means of understanding more deeply the content of existing theories.

It is often claimed that every 'reasonable' *physical* (as opposed to mathematical) model for computation, at least for the deterministic computation of functions from  $\mathbb{Z}$  to  $\mathbb{Z}$ , is equivalent to Turing's. But this is not so; there is no *a priori* reason why physical laws should respect the limitations of the mathematical processes we call 'algorithms' (i.e. the functions  $C(\mathcal{T})$ ). Although I shall not in this paper find it necessary to do so, there is nothing paradoxical or inconsistent in postulating physical systems which compute functions not in  $C(\mathcal{T})$ . There could be experimentally testable theories to that

effect: e.g. consider any recursively enumerable non-recursive set (such as the set of integers representing programs for terminating algorithms on a given Turing machine). In principle, a physical theory might have among its implications that a certain physical device  $\mathcal{F}$  could compute in a specified time whether or not an arbitrary integer in its input belonged to that set. This theory would be experimentally refuted if a more pedestrian Turing-type computer, programmed to enumerate the set, ever disagreed with  $\mathcal{F}$ . (Of course the theory would have to make other predictions as well, otherwise it could never be non-trivially *corroborated*, and its structure would have to be such that its exotic predictions about  $\mathcal{F}$  could not naturally be severed from its other physical content. All this is logically possible.)

Nor, conversely, is it obvious *a priori* that any of the familiar recursive functions is in physical reality computable. The reason why we find it possible to construct, say, electronic calculators, and indeed why we can perform mental arithmetic, cannot be found in mathematics or logic. *The reason is that the laws of physics ‘happen to’ permit the existence of physical models for the operations of arithmetic such as addition, subtraction and multiplication.* If they did not, these familiar operations would be non-computable functions. We might still know of them and invoke them in mathematical proofs (which would presumably be called ‘non-constructive’) but we could not perform them.

If the dynamics of some physical system did depend on a function not in  $C(\mathcal{T})$ , then that system could in principle be used to compute the function. Chaitin (1977) has shown how the truth values of all ‘interesting’ non-Turing decidable propositions of a given formal system might be tabulated very efficiently in the first few significant digits of a single physical constant.

But if they were, it might be argued, we could never know because we could not check the accuracy of the ‘table’ provided by Nature. This is a fallacy. The reason why we are confident that the machines we call calculators do indeed compute the arithmetic functions they claim to compute is not that we can ‘check’ their answers, for this is ultimately a futile process of comparing one machine with another: *Quis custodiet ipsos custodes?* The real reason is that we believe the detailed physical theory that was used in their design. That theory, including its assertion that the abstract functions of arithmetic are realized in Nature, is empirical.

## 2 Quantum computers

Every existing general model of computation is effectively classical. That is, a full specification of its state at any instant is equivalent to the specification of a set of numbers, all of which are in principle measurable. Yet according to quantum theory there exist no physical systems with this property. The fact that classical physics and the classical universal Turing machine do not obey the Church-Turing principle in the strong physical form (1.2) is one motivation for seeking a truly quantum model. The more urgent motivation is, of course, that classical physics is false.

Benioff (1982) has constructed a model for computation within quantum kinematics and dynamics, but it is still effectively classical in the above sense. It is constructed so that at the end of each elementary computational step, no characteristically quantum property of the model —interference, non-separability, or indeterminism— can be detected. Its computations can be perfectly simulated by a Turing machine.

Feynman (1982) went one step closer to a true quantum computer with his ‘universal quantum simulator’. This consists of a lattice of spin systems with nearest-neighbour interactions that are freely specifiable. Although it can surely simulate any system with a finite-dimensional state space (I do not understand why Feynman doubts that it can simulate fermion systems), it is not a computing machine in the sense of this article. ‘Programming’ the simulator consists of endowing it by *fiat* with

the desired dynamical laws, and then placing it in a desired initial state. But the mechanism that allows one to select arbitrary dynamical laws is not modelled. The dynamics of a true ‘computer’ in my sense must be given once and for all, and programming it must consist entirely of preparing it in a suitable *state* (or mixed case).

Albert (1983) has described a quantum mechanical measurement ‘automaton’ and has remarked that its properties on being set to measure itself have no analogue among classical automata. Albert’s automata, though they are not general purpose computing machines, are true quantum computers, members of the general class that I shall study in this section.

In this section I present a general, fully quantum model for computation. I then describe the universal quantum computer  $\mathcal{Q}$ , which is capable of perfectly simulating every finite, realizable physical system. It can simulate ideal closed (zero temperature) systems, including all other instances of quantum computers and quantum simulators, with arbitrarily high but not perfect accuracy. In computing strict functions from  $\mathbb{Z}$  to  $\mathbb{Z}$  it generates precisely the classical recursive functions  $C(\mathcal{T})$  (a manifestation of the correspondence principle). Unlike  $\mathcal{T}$ , it can simulate any finite classical discrete stochastic process perfectly. Furthermore, as we shall see in §3, it has many remarkable and potentially useful capabilities that have no classical analogues.

Like a Turing machine, a model quantum computer  $\mathcal{Q}$ , consists of two components, a finite processor and an infinite memory, of which only a finite portion is ever used. The computation proceeds in steps of fixed duration  $T$ , and during each step only the processor and a finite part of the memory interact, the rest of the memory remaining static.

The processor consists of  $M$  2-state observables

$$\{\hat{n}_i\} \quad (i \in \mathbb{Z}_M) \tag{2.1}$$

where  $\mathbb{Z}_M$  is the set of integers from 0 to  $M - 1$ . The memory consists of an infinite sequence

$$\{\hat{m}_i\} \quad (i \in \mathbb{Z}) \tag{2.2}$$

Of 2-state observables. This corresponds to the infinitely long memory ‘tape’ in a Turing machine. I shall refer to the  $\{\hat{n}_i\}$  collectively as  $\hat{\mathbf{n}}$ , and to the  $\{\hat{m}_i\}$  as  $\hat{\mathbf{m}}$ . Corresponding to Turing’s ‘tape position’ is another observable  $\hat{x}$ , which has the whole of  $\mathbb{Z}$  as its spectrum. The observable  $\hat{x}$  is the ‘address’ number of the currently scanned tape location. Since the ‘tape’ is infinitely long, but will be in motion during computations, it must not be rigid or it could not be made to move ‘by finite means’. A mechanism that moved the tape according to signals transmitted at finite speed between adjacent segments only would satisfy the ‘finite means’ requirement and would be sufficient to implement what follows. Having satisfied ourselves that such a mechanism is possible, we shall not need to model it explicitly. Thus the state of  $\mathcal{Q}$  is a unit vector in the space  $\mathcal{H}$  spanned by the simultaneous eigenvectors

$$|x; \mathbf{n}; \mathbf{m}\rangle \equiv |x; n_0, n_1 \dots n_{M-1}; \dots m_{-1}, m_0, m_1 \dots\rangle \tag{2.3}$$

of  $\hat{x}$ ,  $\hat{\mathbf{n}}$  and  $\hat{\mathbf{m}}$ , labelled by the corresponding eigenvalues  $x$ ,  $\mathbf{n}$  and  $\mathbf{m}$ . I call (2.3) the ‘computational basis states’. It is convenient to take the spectrum of our 2-state observables to be  $\mathbb{Z}_2$ , i.e. the set  $\{0, 1\}$ , rather than  $\{-\frac{1}{2}, +\frac{1}{2}\}$  as is customary in physics. An observable with spectrum  $\{0, 1\}$  has a natural interpretation as a ‘one-bit’ memory element.

The dynamics of  $\mathcal{Q}$  are summarized by a constant unitary operator  $\mathbf{U}$  on  $\mathcal{H}$ .  $\mathbf{U}$  specifies the evolution of any state  $|\psi(t)\rangle \in \mathcal{H}$  (in the Schrödinger picture at time  $t$ ) during a single computation step

$$|\psi(nT)\rangle = \mathbf{U}^n |\psi(0)\rangle \quad (n \in \mathbb{Z}^+) \tag{2.4}$$

$$\mathbf{U}^\dagger \mathbf{U} = \mathbf{U} \mathbf{U}^\dagger = \hat{1}. \quad (2.5)$$

We shall not need to specify the state at times other than non-negative integer multiples of  $T$ . The computation begins at  $t = 0$ . At this time  $\hat{x}$  and  $\hat{\mathbf{n}}$  are prepared with the value zero, the state of a finite number of the  $\hat{\mathbf{m}}$  is prepared as the ‘program’ and ‘input’ in the sense of §1 and the rest are set to zero. Thus

$$\left. \begin{aligned} |\psi(0)\rangle &= \sum_{\mathbf{m}} \lambda_{\mathbf{m}} |0; \mathbf{0}; \mathbf{m}\rangle, \\ \sum_{\mathbf{m}} |\lambda_{\mathbf{m}}|^2 &= 1, \end{aligned} \right\} \quad (2.6)$$

where only a finite number of the  $\lambda_{\mathbf{m}}$  are non-zero and  $\lambda_{\mathbf{m}}$  vanishes whenever an infinite number of the  $\mathbf{m}$  are non-zero.

To satisfy the requirement that  $\mathcal{Q}$  operate ‘by finite means’, the matrix elements of  $\mathbf{U}$  take the following form:

$$\langle x'; \mathbf{n}'; \mathbf{m}' | \mathbf{U} | x; \mathbf{n}; \mathbf{m} \rangle = [\delta_{x'}^{x+1} \mathbf{U}^+(\mathbf{n}', m'_x | \mathbf{n}, m_x) + \delta_{x'}^{x-1} \mathbf{U}^-(\mathbf{n}', m'_x | \mathbf{n}, m_x)] \prod_{y \neq x} \delta_{m_y}^{m_y} \quad (2.7)$$

The continued product on the right ensures that only one memory bit, the  $x$ th, participates in a single computational step. The terms  $\delta_{x'}^{x \pm 1}$  ensure that during each step the tape position  $x$  cannot change by more than one unit, forwards or backwards, or both. The functions  $\mathbf{U}^\pm(\mathbf{n}', m' | \mathbf{n}, m)$ , which represent a dynamical motion depending only on the ‘local’ observables  $\hat{\mathbf{n}}$  and  $\hat{m}_x$ , are arbitrary except for the requirement (2.5) that  $\mathbf{U}$  be unitary. Each choice defines a different quantum computer,  $\mathcal{Q}[\mathbf{U}^+, \mathbf{U}^-]$ .

Turing machines are said to ‘halt’, signalling the end of the computation, when two consecutive states are identical. A ‘valid’ program is one that causes the machine to halt after a finite number of steps. However, (2.4) shows that two consecutive states of a quantum computer  $\mathcal{Q}$  can never be identical after a non-trivial computation. (This is true of any reversible computer.)

Moreover,  $\mathcal{Q}$  must not be observed before the computation has ended since this would, in general, alter its relative state. Therefore, quantum computers need to signal actively that they have halted. One of the processor’s internal bits, say  $\hat{n}_0$ , must be set aside for this purpose. Every valid  $\mathcal{Q}$ -program sets  $n_0$  to 1 when it terminates but does not interact with  $\hat{n}_0$  otherwise. The observable  $\hat{n}_0$  can then be periodically observed from the outside without affecting the operation of  $\mathcal{Q}$ . The analogue of the classical condition for a program to be valid would be that the expectation value of  $\hat{n}_0$  must go to one in a finite time. However, it is physically reasonable to allow a wider class of  $\mathcal{Q}$ -programs. A  $\mathcal{Q}$ -program is valid if the expectation value of its *running time* is finite.

Because of unitarity, the dynamics of  $\mathcal{Q}$ , as of any closed quantum system, are necessarily reversible. Turing machines, on the other hand, undergo irreversible changes during computations, and indeed it was, until recently, widely held that irreversibility is an essential feature of computation. However, Bennett (1973) proved that this is not the case by constructing explicitly a reversible classical model computing machine equivalent to (i.e. generating the same computable function as)  $\mathcal{T}$  (see also Toffoli 1979). (Benioff’s machines are equivalent to Bennett’s but use quantum dynamics.)

Quantum computers  $\mathcal{Q}[\mathbf{U}^+, \mathbf{U}^-]$  equivalent to any reversible Turing machine may be obtained by taking

$$\mathbf{U}^\pm(\mathbf{n}', m' | \mathbf{n}, m) = \frac{1}{2} \delta_{n'}^{\mathbf{A}(\mathbf{n}, m)} \delta_{m'}^{\mathbf{B}(\mathbf{n}, m)} [1 \pm C(\mathbf{n}, m)] \quad (2.8)$$

where  $\mathbf{A}$ ,  $\mathbf{B}$  and  $C$  are functions with ranges  $(\mathbb{Z}_2)^M$ ,  $\mathbb{Z}_2$  and  $\{-1, 1\}$  respectively. Turing machines, in other words, are those quantum computers whose dynamics ensure that they remain in a computational

basis state at the end of each step, given that they start in one. To ensure unitarity it is necessary and sufficient that the mapping

$$\{(\mathbf{n}, m)\} \longleftrightarrow \{(\mathbf{A}(\mathbf{n}, m), \mathbf{B}(\mathbf{n}, m), \mathbf{C}(\mathbf{n}, m))\} \quad (2.9)$$

be bijective. Since the constitutive functions  $\mathbf{A}$ ,  $\mathbf{B}$  and  $\mathbf{C}$  are otherwise arbitrary there must, in particular, exist choices that make  $\mathcal{Q}$  equivalent to a universal Turing machine  $\mathcal{T}$ .

To describe the universal quantum computer  $\mathcal{Q}$  directly in terms of its constitutive transformations  $\mathbf{U}^\pm$  would be possible, but unnecessarily tedious. The properties of  $\mathcal{Q}$  are better defined by resorting to a higher level description, leaving the explicit construction of  $\mathbf{U}^\pm$  as an exercise for the reader. In the following I repeatedly invoke the ‘universal’ property of  $\mathcal{T}$ .

For every recursive function  $f$  there exists a program  $\pi(f)$  for  $\mathcal{T}$  such that when the image of  $\pi(f)$  is followed by the image of any integer  $i$  in the input of  $\mathcal{T}$ ,  $\mathcal{T}$  eventually halts with  $\pi(f)$  and  $i$  themselves followed by the image of  $f(i)$ , with all other bits still (or again) set to zero. That is, for some positive integer  $n$

$$\mathbf{U}^n |0; \mathbf{0}; \pi(f), i, \mathbf{0}\rangle = |0; 1, \mathbf{0}; \pi(f), i, f(i), \mathbf{0}\rangle. \quad (2.10)$$

Here  $\mathbf{0}$  denotes a sequence of zeros, and the zero eigenvalues of  $\hat{m}_i$  ( $i < 0$ ) are not shown explicitly.  $\mathcal{T}$  loses no generality if it is required that every program allocate the memory as an infinite sequence of ‘slots’, each capable of holding an arbitrary integer. (For example, the  $a$ th slot might consist of the bits labelled by successive powers of the  $a$ th prime.) For each recursive function  $f$  and integers  $a, b$  there exists a program  $\pi(f, a, b)$  which computes the function  $f$  on the contents of slot  $a$  and places the result in slot  $b$ , leaving slot  $a$  unchanged. If slot  $b$  does not initially contain zero, reversibility requires that its old value be not overwritten but combined in some reversible way with the value of the function. Thus, omitting explicit mention of everything unnecessary, we may represent the effect of the program  $\pi$  by

$$|\overbrace{\pi(f, 2, 3)}^{\text{slot 1}}, \overbrace{i}^{\text{slot 2}}, \overbrace{j}^{\text{slot 3}}\rangle \rightarrow |\pi(f, 2, 3), i, j \oplus f(i)\rangle, \quad (2.11)$$

where  $\oplus$  is any associative, commutative operator with the properties

$$\left. \begin{array}{l} i \oplus i = 0, \\ i \oplus 0 = i, \end{array} \right\} \quad (2.12)$$

(the exclusive-or function, for example, would be satisfactory). I denote by  $\pi_1 \cdot \pi_2$  the *concatenation* of two programs  $\pi_1$  and  $\pi_2$ , which always exists when  $\pi_1$  and  $\pi_2$  are valid programs;  $\pi_1 \cdot \pi_2$  is a program whose effect is that of  $\pi_1$  followed by  $\pi_2$ .

For any bijective recursive function  $g$  there exists a program  $\phi(g, a)$  whose sole effect is to replace any integer  $i$  in slot  $a$  by  $g(i)$ . The proof is immediate, for if some slot  $b$  initially contains zero,

$$\phi(g, a) = \pi(g, a, b) \cdot \pi(g^{-1}, b, a) \cdot \pi(I, b, a) \cdot \pi(I, a, b). \quad (2.13)$$

Here  $I$  is the ‘perfect measurement’ function (Deutsch 1985)

$$|\pi(I, 2, 3), i, j\rangle \rightarrow |\pi(I, 2, 3), i, j \oplus i\rangle. \quad (2.14)$$

The universal quantum computer  $\mathcal{Q}$  has all the properties of  $\mathcal{T}$  just described, as summarized in (2.10) to (2.14). But  $\mathcal{Q}$  admits a further class of programs which evolve computational basis states

into linear superpositions of each other. All programs for  $\mathcal{Q}$  can be expressed in terms of the ordinary Turing operations and just eight further operations. These are unitary transformations confined to a single two-dimensional Hilbert space  $\mathcal{K}$ , the state space of a single bit. Such transformations form a four (real) parameter family. Let  $\alpha$  be any irrational multiple of  $\pi$ . Then the four transformations

$$\left. \begin{aligned} V_0 &= \begin{pmatrix} \cos \alpha & \sin \alpha \\ -\sin \alpha & \cos \alpha \end{pmatrix}, & V_1 &= \begin{pmatrix} \cos \alpha & i \sin \alpha \\ i \sin \alpha & \cos \alpha \end{pmatrix}, \\ V_2 &= \begin{pmatrix} e^{i\alpha} & 0 \\ 0 & 1 \end{pmatrix}, & V_3 &= \begin{pmatrix} 1 & 0 \\ 0 & e^{i\alpha} \end{pmatrix}, \end{aligned} \right\} \quad (2.15)$$

and their inverses  $V_4, V_5, V_6, V_7$ , generate, under composition, a group dense in the group of all unitary transformations on  $\mathcal{H}$ . It is convenient, though not essential, to add two more generators

$$V_8 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \quad \text{and} \quad V_9 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & i \\ i & 1 \end{pmatrix}, \quad (2.16)$$

which corresponds to  $90^\circ$  ‘spin rotations’. To each generator  $V_i$  there correspond computational basis elements representing programs  $\phi(V_i, a)$ , which perform  $V_i$  upon the least significant bit of the  $a$ th slot. Thus if  $j$  is zero or one, these basis elements evolve according to

$$|\phi(V_i, 2), j\rangle \rightarrow \sum_{k=0}^1 \langle k | V_i | j \rangle |\phi(V_i, 2), k\rangle. \quad (2.17)$$

Composition of the  $V_i$  may be effected by concatenation of the  $\phi(V_i, a)$ . Thus there exist programs that effect upon the state of anyone bit a unitary transformation arbitrarily close to any desired one.

Analogous conclusions hold for the joint state of any finite number  $L$  of specified bits. This is not a trivial observation since such a state is not necessarily a direct product of states confined to the Hilbert spaces of the individual bits, but is in general a linear superposition of such products. However, I shall now sketch a proof of the existence of a program that effects a unitary transformation on  $L$  bits, arbitrarily close to any desired unitary transformation. In what follows, ‘accurate’ means ‘arbitrarily accurate with respect to the inner product norm’. The case  $L = 1$  is trivial. The proof for  $L$  bits is by induction.

First note that the  $(2^L)!$  possible permutations of the  $2^L$  computational basis states of  $L$  bits are all invertible recursive functions, and so can be effected by programs for  $\mathcal{T}$ , and hence for  $\mathcal{Q}$ .

Next we show that it is possible for  $\mathcal{Q}$  to generate  $2^L$ -dimensional unitary transformations diagonal in the computation basis, arbitrarily close to any transformation diagonal in that basis. The  $(L - 1)$ -bit diagonal transformations, which are accurately  $\mathcal{Q}$ -computable by the inductive hypothesis, are generated by certain  $2^L$ -dimensional diagonal unitary matrices whose eigenvalues all have even degeneracy. The permutations of basis states allow  $\mathcal{Q}$  accurately to effect every diagonal unitary transformation with this degeneracy. The closure of this set of degenerate transformations under multiplications is a group of diagonal transformations dense in the group of all  $2^L$ -dimensional diagonal unitary transformations.

Next we show that for each  $L$ -bit state  $|\psi\rangle$  there exists a  $\mathcal{Q}$ -program  $\rho(|\psi\rangle)$  which accurately evolves  $|\psi\rangle$  to the basis state  $|0_L\rangle$  in which all  $L$  bits are zero. Write

$$|\psi\rangle = c_0|0\rangle|\psi_0\rangle + c_1|1\rangle|\psi_1\rangle, \quad (2.18)$$

where  $|\psi_0\rangle$  and  $|\psi_1\rangle$  are states of the  $L - 1$  bits numbered 2 to  $L$ . By the inductive hypothesis there exist  $\mathcal{Q}$ -programs  $\rho_0$  and  $\rho_1$  which accurately evolve  $|\psi_0\rangle$  and  $|\psi_1\rangle$ , respectively, to the  $(L - 1)$ -fold

product  $|0_{L-1}\rangle$ . Therefore there exists a  $\mathcal{Q}$ -program with the following effect. If bit no. 1 is a zero, execute  $\rho_0$  otherwise execute  $\rho_1$ . This converts (2.18) accurately to

$$(c_0|0\rangle + c_1|1\rangle)|0_{L-1}\rangle. \quad (2.19)$$

Then (2.19) can be evolved accurately to  $|0_L\rangle$  by a transformation of bit no. 1.

Finally, an arbitrary  $2^L$ -dimensional transformation  $U$  is accurately effected by successively transforming each eigenvector  $|\psi\rangle$  of  $U$  accurately into  $|0_L\rangle$  (by executing the program  $\rho^{-1}(|\psi\rangle)$ ), then performing a diagonal unitary transformation which accurately multiplies  $|0_L\rangle$  by the eigenvalue (a phase factor) corresponding to  $|\psi\rangle$ , but has arbitrarily little effect on any other computational basis state, and then executing  $\rho(|\psi\rangle)$ .

This establishes the sense in which  $\mathcal{Q}$  is a *universal* quantum computer. It can simulate with arbitrary precision any other quantum computer  $\mathcal{Q}[U^+, U^-]$ . For although a quantum computer has an infinite-dimensional state space, only a finite-dimensional unitary transformation need be effected at every step to simulate its evolution.

### 3 Properties of the universal quantum computer

We have already seen that the universal quantum computer  $\mathcal{Q}$  can perfectly simulate any Turing machine and can simulate with arbitrary precision any quantum computer or simulator. I shall now show how  $\mathcal{Q}$  can simulate various physical systems, real and theoretical, which are beyond the scope of the universal Turing machine  $\mathcal{T}$ .

#### Random numbers and discrete stochastic systems

As is to be expected, there exist programs for  $\mathcal{Q}$  which generate true random numbers. For example, when the program

$$\phi(V_8, 2) \cdot \pi(I, 2, a) \quad (3.1)$$

halts, slot  $a$  contains with probability  $\frac{1}{2}$  either a zero or a one. Iterative programs incorporating (3.1) can generate other probabilities, including any probability that is a recursive real. However, this does not exhaust the abilities of  $\mathcal{Q}$ . So far, all our programs have been, *per se*, classical, though they may cause the ‘output’ part of the memory to enter non-computational basis states. We now encounter our first quantum program. The execution of

$$\frac{1}{\sqrt{2}}|\pi(I, 2, a)\rangle(\cos \theta|0\rangle + \sin \theta|1\rangle) \quad (3.2)$$

yields in slot  $a$ , a bit that is zero with probability  $\cos^2 \theta$ . The whole continuum of states of the form (3.2) are valid programs for  $\mathcal{Q}$ . In particular, valid programs exist with arbitrary irrational probabilities  $\cos^2 \theta$  and  $\sin^2 \theta$ . It follows that every discrete finite stochastic system, whether or not its probability distribution function is  $\mathcal{T}$ -computable, can be perfectly simulated by  $\mathcal{Q}$ . Even if  $\mathcal{T}$  were given access to a ‘hardware random number generator’ (which cannot really exist classically) or a ‘random oracle’ (Bennett 1981) it could not match this. However, it could get arbitrarily close to doing so. But neither  $\mathcal{T}$  nor any classical system whatever, including stochastic ones, can even approximately simulate the next property of  $\mathcal{Q}$ .

## Quantum correlations

The random number generators (3.1) and (3.2) differ slightly from the other programs I have so far considered in that they necessarily produce ‘waste’ output. The bit in slot  $a$  is, strictly speaking, perfectly random only if the contents of slot 2 are hidden from the user and never again participate in computations. The quantum program (3.2) can be used only once to generate a single random bit. If it were re-used the output would contain non-random correlations.

However, in some applications, such correlations are precisely what is required. The state of slots 2 and  $a$  after the execution of (3.1) is the ‘non-separable’ (d’Espagnat 1976) state

$$\frac{1}{\sqrt{2}}(|0\rangle|0\rangle + |1\rangle|1\rangle). \quad (3.3)$$

Consider a pair of programs that swap these slots into an output region of the tape, *one at a time*. That is, if the output is at first blank,

$$\frac{1}{\sqrt{2}}(|0\rangle|0\rangle + |1\rangle|1\rangle)|0\rangle|0\rangle, \quad (3.4)$$

execution of the first program halts with

$$\frac{1}{\sqrt{2}}|0\rangle(|0\rangle|0\rangle + |1\rangle|1\rangle)|0\rangle, \quad (3.5)$$

and, execution of the second program halts with

$$\frac{1}{\sqrt{2}}|0\rangle|0\rangle(|0\rangle|0\rangle + |1\rangle|1\rangle). \quad (3.6)$$

An equivalent program is shown explicitly at the end of §4. Bell’s (1964) theorem tells us that no classical system can reproduce the statistical results of consecutive measurements made on the output slots at times (3.5) and (3.6). (Causing the output to appear in two steps with an opportunity for the user to perform an experiment after each step is sufficient to satisfy the locality requirement in Bell’s theorem.)

The two bits in (3.3) can also be used as ‘keys’ for performing ‘quantum cryptography’ (Bennett *et al.* 1983).

## Perfect simulation of arbitrary finite physical systems

The dynamics of quantum computers, though by construction ‘finite’, are still unphysical in one important respect: the evolution is strictly unitary. However, the third law of thermodynamics (1.3) implies that no realizable physical system can be prepared in a state uncorrelated with systems outside itself, because its entropy would then be zero. Therefore, every realizable physical system interacts with other systems, in certain states. But the effect of its dynamical coupling to systems outside itself cannot be reduced to zero by a finite process because the temperature of the correlation degrees of freedom would then have been reduced to zero. Therefore there can be no realizable way of placing the system in states on which the components of the time evolution operator which mix internal and external degrees of freedom have no effect

A faithful description of a finitely realizable physical system with an  $L$ -dimensional state space  $\mathcal{H}$  cannot therefore be made *via* state vectors in  $\mathcal{H}$  but must use density matrices  $\rho_a^b$ . Indeed, all density matrices are in principle allowed except (thanks to the ‘entropy’ half of the third law (1.3)) pure cases.

The dynamics of such a system are generated not by a unitary operator but by a superscattering matrix  $\$$ :

$$\rho_a{}^b(T) = \sum_{c,d} \$_a{}^{bc}{}_d \rho_c{}^d(0). \quad (3.7)$$

It is worth stressing that I am not advocating non-unitary dynamics for the universe as a whole, which would be a heresy contrary to quantum theory. Equation (3.7) is, of course, merely the projection into  $\mathcal{H}$  of unitary evolution in a higher state space  $\mathcal{H} \times \mathcal{H}'$ , where  $\mathcal{H}'$  represents as much of the rest of the universe as necessary. Roughly speaking (the systems are far from equilibrium)  $\mathcal{H}'$  plays the role of a ‘heat bath’.

Thus the general superscattering operator has the form

$$\$_a{}^{bc}{}_d = \sum_{e',f',g'} \mathbf{U}_{ae'}{}^{cf'} \mathbf{U}^{bc'}{}_{dg'} \bar{\rho}_{f'}{}^{g'}, \quad (3.8)$$

where  $\mathbf{U}_{ab'}{}^{cd'}$  is a unitary operator on  $\mathcal{H} \times \mathcal{H}'$ , that is

$$\sum_{c,d'} \mathbf{U}_{ab'}{}^{cd'} \mathbf{U}^{ef'}{}_{cd'} = \delta_a{}^e \delta_{b'}{}^{f'}, \quad (3.9)$$

which does not decompose into a product of operators on  $\mathcal{H}$  and  $\mathcal{H}'$ . (Raising and lowering of indices denotes complex conjugation.) The term  $\bar{\rho}_{a'}{}^{b'}$  has an approximate interpretation as the initial density matrix of the ‘heat bath’, which would be strictly true if the system, the heat bath, and the entity preparing the system in its initial state were all uncorrelated initially. Let us rewrite (3.8) in the  $\mathcal{H}'$ -basis in which  $\bar{\rho}$  is diagonal :

$$\begin{aligned} \$_a{}^{bc}{}_d &= \sum_{e',f'} P_{f'} \mathbf{U}_{ae'}{}^{cf'} \mathbf{U}^{be'}{}_{df'}, \\ \sum_{a'} P_{a'} &= 1, \end{aligned} \quad (3.10)$$

where the probabilities  $P_{a'}$  are the eigenvalues of  $\bar{\rho}$ . The set  $\mathbb{G}$  of all superscattering matrices (3.8) or (3.10) lies in a subspace  $\mathcal{J}$  of  $\mathcal{H} \times \mathcal{H}^* \times \mathcal{H}^* \times \mathcal{H}$ , namely the subspace whose elements satisfy

$$\sum_a \$_a{}^{bc}{}_c = \delta_b{}^c. \quad (3.11)$$

Every element of  $\mathbb{G}$  satisfies the constraints

$$0 \leq \sum_{a,b,c,d} \rho^{(1)}_b \$_a{}^{bc}{}_d \rho^{(2)}_c{}^d \leq 1 \quad (3.12)$$

for arbitrary density matrices  $\rho^{(l)}$  and  $\rho^{(2)}$ .

The inequality on the left in (3.12) can be an equality only if the states of  $\mathcal{H}$  form disjoint subsets with strictly zero probability so that thermal noise can effect a transition between them. This is impossible unless there are superselection rules forbidding such transitions, a possibility that we lose no generality by excluding because only one superselected sector at a time can be realized as a physical system. The inequality on the right becomes an equality precisely in the unitary case

$$\$_a{}^{bc}{}_d = \mathbf{U}_a{}^c \mathbf{U}_d{}^b, \quad (3.13)$$

which is unphysical because it represents perfectly non-dissipative evolution. Thus the set of physically realizable elements of  $\mathbb{G}$  is an open set in  $\mathcal{J}$ . Moreover, for any  $\$^{(1)}$  and  $\$^{(2)}$  that are  $\mathcal{Q}$ -computable the convex linear combination

$$p_1\$^{(1)} + p_2\$^{(2)}, \quad (3.14)$$

where  $p_1$  and  $p_2$  are arbitrary probabilities, is also computable, thanks to the random number generator (3.2). By computing unitary transformations as in (3.10), every element of a certain countable dense subset of  $\mathbb{G}$  can be computed. But every point in any open region of a finite-dimensional vector space can be represented as a finite convex linear combination of elements of any dense subset of that space. It follows that  $\mathcal{Q}$ , can perfectly simulate any physical system with a finite-dimensional state space. Therefore quantum theory is compatible with the Church-Turing principle (1.2).

The question whether all finite systems in the physical universe can likewise be simulated by  $\mathcal{Q}$ , — i.e. whether (1.2) is satisfied in Nature — must remain open until the state space and dynamics of the universe are understood better. What little is known seems to bear out the principle. If the theory of the thermodynamics of black holes is trustworthy, no system enclosed by a surface with an appropriately defined area  $A$  can have more than a finite number (Bekenstein 1981)

$$N(A) = \exp(Ac^3/4\hbar G) \quad (3.15)$$

of distinguishable accessible states ( $\hbar$  is the Planck reduced constant,  $G$  is the gravitational constant and  $c$  is the speed of light). That is, in a suitable basis the system can be perfectly described by using an  $N(A)$ -dimensional state space, and hence perfectly simulated by  $\mathcal{Q}$ .

### Parallel processing on a serial computer

Quantum theory is a theory of parallel interfering universes. There are circumstances under which different computations performed in different universes can be combined by  $\mathcal{Q}$  giving it a limited capacity for parallel processing. Consider the quantum program

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N |\pi(f, 2, 3), i, 0\rangle, \quad (3.16)$$

which instructs  $\mathcal{Q}$  in each of  $N$  universes to compute  $f(i)$ , for  $i$  from 1 to  $N$ . Linearity and (2.11) imply that after executing (3.16)  $\mathcal{Q}$  halts in the state

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N |\pi(f, 2, 3), i, f(i)\rangle. \quad (3.17)$$

Although this computation requires exactly the same time, memory space and hardware as (2.11), the state (3.17) contains the results of an arbitrarily large number  $N$  of separate computations. Unfortunately, at most one of these results is accessible in each universe. If (3.16) is executed many times, the mean time required to compute all  $N$  values  $f(i)$ , which I shall refer to collectively as  $\mathbf{f}$ , is at least that required for (2.11) to compute all of them serially. I shall now show that the expectation value of the time to compute any non-trivial  $N$ -fold parallelizable function  $G(\mathbf{f})$  of all  $N$  values  $\mathbf{f}$  via quantum parallelism such as (3.16) cannot be less than the time required to compute it serially via (2.11).

For simplicity assume that  $\tau$ , the running time of (2.11), is independent of  $i$  and that the time taken to combine all the  $\mathbf{f}$  to form  $G(\mathbf{f})$  is negligible compared with  $\tau$ . Now suppose that there exists a

program  $\zeta$ , which for any function  $f$  extracts the value of  $G(\mathbf{f})$  from (3.17) in a negligible time and with probability  $|\beta|^2$ . That is,  $\zeta$  has the effect

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N |i, f(i)\rangle \rightarrow \beta|0, G(\mathbf{f})\rangle + \sqrt{1 - |\beta|^2}|1\rangle|\lambda(\mathbf{f})\rangle, \quad (3.18)$$

where the states  $|\lambda(\mathbf{f})\rangle$  contain no information about  $G(\mathbf{f})$ . Then the first slot could be measured. If it contained zero, the second slot would contain  $G(\mathbf{f})$ . Otherwise the information in (3.17) would have been lost and it would have to be recomputed. Unitarity implies

$$\frac{1}{n} \sum_{i=1}^N \delta(f(i), g(i)) = |\beta|^2 \delta(G(\mathbf{f}), G(\mathbf{g})) + (1 - |\beta|^2) \langle \lambda(\mathbf{f}) | \lambda(\mathbf{g}) \rangle \quad (3.19)$$

for any functions  $g(i)$  and  $f(i)$ .

If  $G(\mathbf{f})$  is not a constant function then for each function  $f(i)$  there exists another function  $g(i)$  such that  $G(\mathbf{g}) \neq G(\mathbf{f})$ , but  $g(i) = f(i)$  for all but one value of  $i$  between 1 and  $N$ . For this choice

$$1 - \frac{1}{N} = (1 - |\beta|^2) \langle \lambda(\mathbf{f}) | \lambda(\mathbf{g}) \rangle, \quad (3.20)$$

whence it follows that  $|\beta|^2 < 1/N$ . Thus the mean time to compute  $G(\mathbf{f})$  must be at least  $\tau/|\beta|^2 = N\tau$ . This establishes that quantum parallelism cannot be used to improve the mean running time of parallelizable algorithms. As an example of quantum parallelism for  $N = 2$ , let

$$G(\mathbf{f}) \equiv f(0) \oplus f(1), \quad (3.21)$$

(see equations (2.12)). Then the state (3.17) following the quantum parallel computation has

$$\frac{1}{\sqrt{2}}(|0, f(0)\rangle + |1, f(1)\rangle) \quad (3.22)$$

as a factor. A suitable program  $\zeta$  to ‘decode’ this is one that effects a measurement of any non-degenerate observable with eigenstates

$$\begin{aligned} |\text{zero}\rangle &\equiv \frac{1}{2}(|0, 0\rangle - |0, 1\rangle + |1, 0\rangle - |1, 1\rangle), \\ |\text{one}\rangle &\equiv \frac{1}{2}(|0, 0\rangle - |0, 1\rangle - |1, 0\rangle + |1, 1\rangle), \\ |\text{fail}\rangle &\equiv \frac{1}{2}(|0, 0\rangle + |0, 1\rangle + |1, 0\rangle + |1, 1\rangle), \\ |\text{error}\rangle &\equiv \frac{1}{2}(|0, 0\rangle + |0, 1\rangle - |1, 0\rangle - |1, 1\rangle). \end{aligned} \quad \left. \right\} \quad (3.23)$$

Such an observable exists, since the states (3.23) form an orthonormal set. Furthermore, the measurement can be made in a fixed time independent of the execution time of the algorithm computing  $f$ . If the outcome of the measurement is ‘zero’ (i.e. the eigenvalue corresponding to the state  $|\text{zero}\rangle$ ) or ‘one’ then it can be inferred that  $f(0) \oplus f(1)$  is zero or one respectively. Whatever the form of the function  $f$ , there will be a probability 1/2 that the outcome will be ‘fail’, in which case nothing can be inferred about the value of  $f(0) \oplus f(1)$ . The probability of the outcome ‘error’ can be made arbitrarily small with a computational effort independent of the nature of  $f$ .

In this example the bound  $N\tau$  for the running time has been attained. However, for  $N > 2$  I have been unable to construct examples where the mean running time is less than  $(N^2 - 2N + 2)\tau$ , and I conjecture that this is the optimal lower bound. Also, although there exist non-trivial examples of

quantum parallelizable algorithms for all  $N$ , when  $N > 2$  there are none for which the function  $G(\mathbf{f})$  has the set of all  $2^N$  possible graphs of  $f$  as its domain.

In practical computing problems, especially in real time applications, one may not be concerned with minimizing specifically the *mean* running time of a program: often it is required that the minimum or maximum time or some more complicated measure be minimized. In such cases quantum parallelism may come into its own. I shall give two examples.

(1) Suppose that (3.17) is a program to estimate tomorrow's Stock Exchange movements given today's, and  $G(\mathbf{f})$  specifies the best investment strategy. If  $\tau$  were one day and  $N = 2$ , the classical version of this program would take two days to run and would therefore be useless. If the quantum version was executed every day, then on one day in two on average slot 1 would contain the measured value '1', indicating a failure. On such days one would make no investment. But with equal average frequency a zero would appear, indicating that slot 2 contained the correct value of the investment strategy  $G(\mathbf{f})$ .  $G(\mathbf{f})$ , which incorporates the result of two classical processor-days of computation, would on such occasions have been performed by one processor in one day.

One physical way of describing this effect is that when the subtasks of an  $N$ -fold parallel task are delegated to  $N^2 - 2N + 2$  universes, at most one of them can acquire the overall result.

(2) Now consider the problem of the design of parallel information-processing systems which are subject to noise. For example, suppose that it is required, within a fixed time  $\tau$ , to compute a certain  $N$ -fold parallelizable function  $G(\mathbf{f})$ .  $NR$  processors are available, each of which may fail for reasons of thermal noise, etc. with probability  $p$ . For simplicity assume that such a hardware error can be reliably detected. The problem is to minimize the overall failure rate  $q$ . 'Classically' (i.e. without using quantum parallelism) one minimizes  $q$  by means of an  $R$ -fold redundancy:  $R$  processors are instructed to perform each of the  $N$  parallel subtasks. The machine as a whole will therefore fail to compute the result in time only when all  $R$  processors assigned to anyone subtask fail, and this occurs with probability

$$q_{\text{classical}} = 1 - (1 - p^R)^N. \quad (3.24)$$

Using quantum parallelism, however, each of the  $NR$  available processors may be given all  $N$  tasks. Each is subject to two independent causes of failure, (i) the probability  $p$  that it will fail for hardware reasons, and (ii) the probability, which as I have indicated will for certain  $G(\mathbf{f})$  be  $1 - 1/(N^2 - 2N + 2)$ , that it will end up in a different universe from the answer. It takes only one of the  $NR$  processors to succeed, so the failure rate is

$$q_{\text{quantum}} = \left[ 1 - \frac{1 - p}{N^2 - 2N + 2} \right]^{NR} \quad (3.25)$$

a number which, for suitable values of  $p$ ,  $N$  and  $R$ , can be smaller than (3.24).

## Faster computers

One day it will become technologically possible to build quantum computers, perhaps using flux quanta (Likharev 1982; Leggett 1985) as the fundamental components. It is to be expected that such computers could operate at effective computational speeds in excess of Turing-type machines built with the same technology. This may seem surprising since I have established that no recursive function can be computed by  $\mathcal{Q}$  on average more rapidly with the help of quantum programs than without. However, the idealizations in  $\mathcal{Q}$  take no account of the purely technological fact that it is always easier in practice to prepare a very large number of identical systems in the same state than to

prepare each in a different state. It will therefore be possible to use a far higher degree of redundancy  $R$  for parallel quantum programs than for classical ones running on the same basic hardware.

### Interpretational implications

I have described elsewhere (Deutsch 1985; cf. also Albert 1983) how it would be possible to make a crucial experimental test of the Everett ('many-universes') interpretation of quantum theory by using a quantum computer (thus contradicting the widely held belief that it is not experimentally distinguishable from other interpretations). However, the performance of such experiments must await both the construction of quantum computers and the development of true artificial intelligence programs. In explaining the operation of quantum computers I have, where necessary, assumed Everett's ontology. Of course the explanations could always be 'translated' into the conventional interpretation, but not without entirely losing their explanatory power. Suppose, for example, a quantum computer were programmed as in the Stock Exchange problem described. Each day it is given different data. The Everett interpretation explains well how the computer's behaviour follows from its having delegated subtasks to copies of itself in other universes. On the days when the computer succeeds in performing two processor-days of computation, how would the conventional interpretations explain the presence of the correct answer? *Where was it computed?*

## 4 Further connections between physics and computer science

### Quantum complexity theory

Complexity theory has been mainly concerned with constraints upon the computation of functions: which functions can be computed, how fast, and with use of how much memory. With quantum computers, as with classical stochastic computers, one must also ask 'and with what probability?'. We have seen that the minimum computation time for certain tasks can be lower for  $\mathcal{Q}$  than for  $\mathcal{T}$ . Complexity theory for  $\mathcal{Q}$  deserves further investigation.

The less immediately applicable but potentially more important application of complexity theory has been in the attempt to understand the spontaneous growth of complexity in physical systems, for example the evolution of life, and the growth of knowledge in human minds. Bennett (1983) reviewed several different measures of complexity (or 'depth', or 'knowledge') that have been proposed. Most suffer from the fatal disadvantage that they assign a high 'complexity' to a purely random state. Thus they do not distinguish true knowledge from mere information content. Bennett has overcome this problem. His 'logical depth' is roughly the running time of the shortest  $\mathcal{T}$ -program that would compute a given state  $\psi$  from a blank input. Logical depth is at a minimum for random states. Its intuitive physical justification is that the 'likeliest explanation' why a physical system might be found to be in the state  $\psi$  is that  $\psi$  was indeed 'computed' from that shortest  $\mathcal{T}$ -program. In biological terminology, logical depth measures the amount of evolution that was needed to evolve  $\psi$  from the simplest possible precursors.

At first sight Bennett's construction seems to lose this physical justification when it is extended beyond the strictly deterministic physics of Turing machines. In physical reality most random states are not generated by 'long programs' (i.e. precursors whose complexity is near to their own), but by short programs relying on indeterministic hardware. However, there is a quantum analogue of Bennett's idea which solves this problem. Let us define the Q-logical depth of a quantum state as the running time of the shortest  $\mathcal{Q}$ -program that would generate the state from a blank input (or, perhaps,

as Bennett would have it, the harmonic mean of the running times of all such programs). Random numbers can be rapidly generated by *short*  $\mathcal{Q}$ -programs.

Notice that the Q-logical depth is not even in principle an observable, because it contains information about all universes at once. But this makes sense physically: the Q-logical depth is a good measure of knowledge in that it gives weight only to complexity that is present in all universes, and can therefore be assumed to have been put there ‘deliberately’ by a deep process. Observationally complex states that are different in different universes are not truly deep but just random. Since the Q-logical depth is a property of the quantum *state* (vector), a quantum subsystem need not necessarily have a well defined Q-logical depth (though often it will to a good degree of approximation). This is again to be expected since the knowledge in a system may reside entirely in its correlations with other systems. A spectacular example of this is quantum cryptography.

### Connections between the Church-Turing principle and other parts of physics

We have seen that quantum theory obeys the strong form (1.2) of the Church-Turing principle only on the assumption that the third law of thermodynamics (1.3) is true. This relation is probably better understood by considering the Church-Turing principle as more fundamental and deriving the third law from it and quantum theory.

The fact that classical physics does not obey (1.2) tempts one to go further. Some of the features that distinguish quantum theory from classical physics (for example the discreteness of observables?) can evidently be derived from (1.2) and the laws of thermodynamics alone. The new principle has therefore given us at least part of the solution to Wheeler’s problem ‘Why did quantum theory have to be?’ (see, for example, Wheeler 1985).

Various ‘arrows of time’ that exist in different areas of physics have by now been connected and shown to be different manifestations of the same effect. But, contrary to what is often asserted, the ‘psychological’ or ‘epistemological’ arrow of time is an exception. Before Bennett (1973) it could be maintained that computation is intrinsically irreversible, and since psychological processes such as the growth of knowledge are computations, the psychological arrow of time is necessarily aligned with the direction in which entropy increases. This view is now untenable, the alleged connection fallacious.

One way of reincorporating the psychological arrow of time into physics is to postulate another new principle of Nature which refers directly to the Q-logical depth. It seems reasonable to assert, for example, that the Q-logical depth of the universe is at a minimum initially. More optimistically the new principle might require the Q-logical depth to be non-decreasing. It is perhaps not unreasonable to hope that the second law of thermodynamics might be derivable from a constraint of this sort on the Q-logical depth. This would establish a valid connection between the psychological (or epistemological, or evolutionary) and thermodynamic ‘arrows of time’.

### Programming physics

To view the Church-Turing hypothesis as a physical principle does not merely make computer science a branch of physics. It also makes part of experimental physics into a branch of computer science.

The existence of a universal quantum computer  $\mathcal{Q}$  implies that there exists a program for each physical process. In particular,  $\mathcal{Q}$  can perform any physical experiment. In some cases (for example measurement of coupling constants or the form of interactions) this is not useful because the result must be known to write the program. But, for example, when testing quantum theory itself, every

experiment is genuinely just the running of a  $\mathcal{Q}$ -program. The execution on  $\mathcal{Q}$  of the following ALGOL 68 program is a performance of the Einstein-Podolski-Rosen experiment:

```

begin
  int  $n = 8 * \text{random};$           % random integer from 0 to 7 %
  bool  $x, y;$                   % bools are 2-state memory elements %
   $x := y := \text{false};$           % an irreversible preparation %
   $V(8, y);$                    % see equation (2.15) %
   $x \text{ eorab } y;$             % perfect measurement (2.14) %
  if  $V(n, y) \neq$               % measure y in random direction %
     $V(n, x)$                   % and x in the parallel direction %
  then  $\text{print}(\text{"Quantum theory refuted."})$ 
  else  $\text{print}(\text{"Quantum theory corroborated."})$ 
fi
end

```

Quantum computers raise interesting problems for the design of programming languages, which I shall not go into here. From what I have said, programs exist that would (in order of increasing difficulty) test the Bell inequality, test the linearity of quantum dynamics, and test the Everett interpretation. I leave it to the reader to write them.

I wish to thank Dr C. H. Bennett for pointing out to me that the Church-Turing hypothesis has physical significance, C. Penrose and K. Wolf for interesting discussions about quantum computers, and Professor R. Penrose, F.R.S., for reading an earlier draft of the article and suggesting many improvements.

This work was supported in part by N.S.F. grant no. PHY 8205717.

## References

- [1] Albert, D. Z. 1983 *Phys. Lett. A* **98**, 249.
- [2] Bekenstein, J. D. 1973 *Phys. Rev. D* **7**, 2333.
- [3] Bekenstein, J. D. 1981 *Phys. Rev. D* **23**, 287.
- [4] Bell, J. S. 1964 *Physica* **1**, 195.
- [5] Benioff, P. A. 1982 *Int. J. theor. Phys.* **21**, 177.
- [6] Bennett, C. H. 1973 *IBM Jl Res. Dev.* **17**, 525.
- [7] Bennett, C. H. 1981 *SIAM Jl Comput.* **10**, 96.
- [8] Bennett, C. H. 1983 On various measures of complexity, especially ‘logical depth’. Lecture at Aspen. IBM Report.
- [9] Bennett, C. H., Brassard, G., Breidbart, S. & Wiesner, S. 1983 Advances in cryptography. In *Proceedings of Crypto 82*. New York: Plenum.

- [10] Chaitin, G. J. 1977 *IBM Jl Res. Dev.* **21**, 350.
- [11] Church, J. 1936 *Am. J. Math.* **58**, 435.
- [12] Deutsch, D. 1985 *Int. J. theor. Phys.* **24**, 1.
- [13] d'Espagnat, B. 1976 *Conceptual foundations of quantum mechanics* (second edn). Reading, Massachusetts: W. A. Benjamin.
- [14] Feynman, R. P. 1982 *Int. J. theor. Phys.* **21**, 467.
- [15] Gandy, R. 1980 In *The Kleene symposium* (ed. J. Barwise, H. J. Keisler & K. Kunen), pp. 123-148. Amsterdam: North Holland.
- [16] Hofstadter, D. R. J 1979 *Gödel, Escher, Bach: an eternal golden braid*. New York: Random House.
- [17] Leggett, A. J. 1985 In *Quantum discussions, proceedings of the Oxford quantum gravity conference 1984* (ed. R. Penrose & C. Isham). Oxford University press.
- [18] Likharev, K. K. 1982 *Int. J. theor. Phys.* **21**, 311.
- [19] Popper, K. R. 1959 *The logic of scientific discovery*. London: Hutchinson.
- [20] Toffoli, T. J. 1979 *J. Comput. Syst. Sci.* **15**, 213.
- [21] Turing, A. M. 1936 *Proc. Lond. math. Soc. Ser. 2*, **442**, 230.
- [22] Wheeler, J. A. 1985 In *NATO Advanced Study Institute Workshop on Frontiers of Nonequilibrium Physics 1984*. New York: Plenum.



## New Variables for Classical and Quantum Gravity

Abhay Ashtekar

*Physics Department, Syracuse University, Syracuse, New York 13244, and Institute for Theoretical Physics,  
University of California, Santa Barbara, Santa Barbara, California 93106*

(Received 18 December 1985; revised manuscript received 29 August 1986)

A Hamiltonian formulation of general relativity based on certain spinorial variables is introduced. These variables simplify the constraints of general relativity considerably and enable one to imbed the constraint surface in the phase space of Einstein's theory into that of Yang-Mills theory. The imbedding suggests new ways of attacking a number of problems in both classical and quantum gravity. Some illustrative applications are discussed.

PACS numbers: 04.60.+n, 04.20.Fy

Attempts at constructing perturbative quantum gravity have been unsuccessful. It is now generally believed that the problem lies in the basic assumption of the perturbation theory that the true space-time structure can be well approximated by a classical background geometry even below the Planck scale. From this standpoint, there is little hope in retaining the general perturbative framework and simply changing, e.g., the form of the Einstein Lagrangean by adding higher-derivative terms or supersymmetric matter. A more promising direction is to face the problem nonperturbatively. For, as has been emphasized by J. Klauder over the years, quantum gravity may well exist as an exact theory in spite of perturbative nonrenormalizability. The canonical quantization scheme provides a natural avenue in this direction since it does not require the fixation of a classical background geometry. Furthermore, the fact that the Hamiltonian structure of general relativity has certain essentially non-perturbative features—in the exact theory, the Hamiltonian is essentially given by the constraints, while the two decouple in any order in perturbation theory—suggests that qualitatively new results may arise from exact canonical quantization.

Over the years, the main obstacle to the canonical quantization program has come from the fact that the constraint equations have a complicated, nonpolynomial dependence on the traditional canonically conjugate variables. The purpose of this Letter is to report the existence of new variables in terms of which the constraints simplify considerably and to point out that the use of these variables provides new, nonperturbative approaches to problems in both classical and quantum gravity.

Let us begin with some mathematical preliminaries. Fix a three-manifold  $\Sigma$  and consider on it, in addition to tensor fields,  $T^{a\dots b}_{c\dots d}$ , fields with “internal” SU(2) indices,  $\lambda^A, \mu_A, \dots$ . The SU(2) structure provides us with volume forms,  $\epsilon^{AB}$  and  $\epsilon_{AB}$ , on internal indices, satisfying  $\epsilon^{AB}\epsilon_{AD} = \delta^B_D$ . We shall use them to raise and lower these indices:  $\lambda^A := \epsilon^{AB}\lambda_B$  and  $\mu_A := \mu^B\epsilon_{BA}$ . Now, given any isomorphism  $\sigma_B^A$  between tangent vectors  $V^a$  and second-rank, trace-free, Hermitian fields  $V^A_B$ , the identification  $V^a \equiv -\sigma_B^AV^B_A$  solders the internal indices to the

tangent space of  $\Sigma$  and makes them SU(2) spinor indices. Furthermore, each soldering form,  $\sigma$ , is a “square root” of a (positive definite) metric  $q_{ab}$  on  $\Sigma$ ,

$$q_{ab} := \sigma_a^{AB}\sigma_b^{MN}\epsilon_{AM}\epsilon_{BN} = -\text{Tr}\sigma_a\sigma_b, \quad (1)$$

and singles out a unique torsion-free connection  $D$  on tensor and spinor fields on  $\Sigma$  satisfying

$$D_a\epsilon_{AB} = 0, \quad D_a\sigma^{bA}_B = 0. \quad (2)$$

The configuration space  $\mathcal{C}$  for general relativity is to be the space of all (suitably) regular and, if  $\Sigma$  is noncompact, asymptotically well-behaved soldering forms  $\sigma^{aA}_B$ . The phase space  $\Gamma$  is the cotangent bundle over  $\mathcal{C}$ . Thus, a point of  $\Gamma$  consists of a pair  $(\sigma^{aA}_B, M_m^M{}_N)$ , where  $M$ , a density of weight 1, is the momentum canonically conjugate to  $\sigma$ . The canonical variables  $(q_{ab}, p^{ab})$  of the traditional Hamiltonian formulation<sup>1</sup> are now to be regarded as “derived” quantities:  $q_{ab}$  is given by (1) and  $p^{ab}$  by

$$p^{ab} := -\text{Tr}M_m\sigma^{(a}q^{b)m} \equiv M^{(ab)}. \quad (3)$$

As usual, not all points of  $\Gamma$  are accessible to the physical gravitational field: There are constraints. First, we have the familiar constraints

$$C^b(\sigma, M) := D_ap^{ab} = 0, \quad (4)$$

$$C(\sigma, M) := (p^{ab}p_{ab} - \frac{1}{2}p^2) - \frac{(\det q)}{G^2}R = 0, \quad (5)$$

where  $R$  is the scalar curvature of  $q_{ab}$ . However, since we have enlarged the configuration space from the space of the six-component fields  $q_{ab}$  to that of the nine-component fields  $\sigma^{aA}_B$ , we have three new constraints:

$$M^{[ab]} = 0. \quad (6)$$

The canonical transformations generated by these constraints cause SU(2) rotations on the internal indices of  $\sigma$  and  $M$ .

The above framework is equivalent to the familiar triad formalisms. The key new step is the introduction of new variables on  $\Gamma$ . Given any point  $(\sigma, M)$  of  $\Gamma$ , introduce

two connections  $\pm \mathcal{D}$  on  $\Sigma$ :

$$\pm \mathcal{D}_a a_{bM} = D_a a_{bM} \pm (i/\sqrt{2}) \Pi_{aM}^N a_{bN}, \quad (7)$$

where  $\Pi_{aM}^N$  is given by<sup>2</sup>

$$\begin{aligned} \Pi_{aM}^N &= G(\det q)^{-1/2} (M_{aM}^N - \frac{1}{2} \sigma_{aM}^N \sigma_{AB}^B M_B^{AB}), \quad (8) \\ &= G(\det q)^{-1/2} (M_{aM}^N - \frac{1}{2} \sigma_{aM}^N \sigma_{AB}^B M_B^{AB}). \end{aligned}$$

The use of these connections simplifies the structure of constraints (4) and (5) considerably. To see this, introduce connection one-forms  $\pm A_a$  and curvature two-forms  $\pm F_{ab}$  associated with  $\pm \mathcal{D}$ :

$$\pm \mathcal{D}_a a_M = \partial_a a_M + G \pm A_{aM}^N a_N, \quad (9)$$

$$2 \pm \mathcal{D}_{[a} \pm \mathcal{D}_{b]} a_M = :G \pm F_{abM}^N a_N, \quad (10)$$

where  $\partial$  is a fixed ( $c$  number) connection, also satisfying  $\partial_a \epsilon_{AB} = 0$ . Then (6) can be rewritten as

$$\pm \mathcal{D}_a \sigma^a_{AB} = 0, \quad (6')$$

and, modulo (6), Eqs. (4) and (5) can be recast as

$$\text{Tr} \sigma^a \pm F_{ab} = 0, \quad (4')$$

$$\text{Tr} \sigma^a \sigma^b \pm F_{ab} = 0. \quad (5')$$

(Throughout,  $\pm$  stands for  $+$  or  $-$ ; we can use either  $+A_a$  or  $-A_a$ .) Constraints (4')–(6') are closed under the Poisson bracket and preserved by dynamics.

Note that the form of constraints (4')–(6') is simpler than that of (4)–(6) in at least two respects. First, (4')–(6') are at most quadratic in each of the new variables ( $\sigma^a, \pm A_a$ ) while (4)–(6) involve nonpolynomial functions of  $q_{ab}$ . Second, if one were to regard  $\pm A_a$  as the new configuration variable and  $\sigma^a$  as the “momentum,” (5') involves only a “kinetic” term, quadratic in the new momenta, and is therefore structurally similar to the strong-coupling limit of (5) in which the “potential” term,  $R$ , is neglected. [This came about because  $\pm A_a$  knows both about  $p^{ab}$  and (the connection of)  $q_{ab}$ .] These features lead to new avenues especially in the quantum theory.

What is the physical interpretation of  $\pm A_a$ ? Consider a solution of  $g_{ab}$  of Einstein's equation obtained from initial data  $(\sigma^a, M_a)$  [satisfying (4)–(6)]. Then one has the following:  $\pm \mathcal{D}$  are the restrictions to  $\Sigma$  of the four-dimensional spin-connection  $\nabla$  on (un)primed  $SL(2, C)$  spinors (e.g.,  ${}^+ \mathcal{D}_a \lambda_M = q_a^b \nabla_b \lambda_m$ ), and  $\text{Tr} \pm F_{ab} \sigma^c \epsilon^{abd} = (\sqrt{2}/G)(E^{cd} \mp iB^{cd})$ , where  $E^{cd}$  and  $B^{cd}$  are the electric and magnetic parts, relative to  $\Sigma$ , of the Weyl tensor of  $g_{ab}$ . Thus,  $\pm A_a$  is a potential for the (anti-)self-dual part of the Weyl tensor. This fact leads to an interesting application: One can obtain (complex Lorentzian or, with minor convention changes, real Euclidean) self-dual solutions to Einstein's equation by simply setting  $\pm A_a = 0$ . This *Ansatz* trivializes (4') and (5') and simplifies (6') as well as the evolution equations (which, in-

cidentally, automatically preserve the *Ansatz*). The resulting system of equations provides a new, simple, and convenient characterization of self-dual solutions.<sup>3</sup> Traditionally,  $H$ -space and twistor techniques have been used to study these solutions.<sup>4</sup> The new variables serve to bridge these techniques to the Hamiltonian methods. Other applications, to classical relativity, include the following: analysis of gravitational perturbations; interesting, exact solutions to constraint equations; understanding and generalizing of the results<sup>5</sup> on the relation between certain classes of solutions to Einstein's and Yang-Mills equations; and the use and role of hypersurface twistors in general relativity.

On the phase space  $\Gamma$ , the new variables have several interesting properties. Each of  $\{ {}^+ A_a \}$ ,  $\{ {}^- A_a \}$ , and  $\{ \tilde{\sigma}^a \equiv (\det q)^{1/2} \sigma^a \}$  forms a complete set of commuting variables with respect to the natural Poisson bracket on  $\Gamma$ . Furthermore,  $\pm A_a$  is “conjugate” to  $\tilde{\sigma}^a$  in the sense that

$$\begin{aligned} &\{ \pm A_m^{MN}(x), \tilde{\sigma}^a_{AB}(y) \}_{\text{P.B.}} \\ &= \pm (i/\sqrt{2}) \delta_m^a \delta_A^M \delta_B^B \delta(x, y). \quad (11) \end{aligned}$$

[The factor of  $G$  in (9) ensures that  $(A) \cdot (\tilde{\sigma})$  has dimensions of action.] I shall therefore use  $\tilde{\sigma}^a$  and  $\pm A_a$  as the basic variables. [Note that one can replace  $\sigma^a$  by  $\tilde{\sigma}^a$  in the constraints (4')–(6') free of charge.] These properties suggest identifications with certain variables that are featured in the Yang-Mills theory.  $\pm A_a$  is the connection one-form;  $\pm B^a := \epsilon^{abc} \pm F_{bc}$ , its magnetic field; and  $\tilde{\sigma}^a$ , the analog of the electric field  $E^a$ . In terms of these Yang-Mills variables, the constraints become

$$\pm \mathcal{D} \cdot \mathbf{E} = 0, \quad (6'')$$

$$\text{Tr} \mathbf{E} \times \mathbf{B} = 0, \quad (4'')$$

$$\text{Tr} \mathbf{E} \cdot (\mathbf{E} \times \mathbf{B}) = 0. \quad (5'')$$

Thus, every initial datum (satisfying constraints)  $(\sigma, M)$  for Einstein's equation yields initial data  $(\mathbf{A}, \mathbf{E})$  for Yang-Mills equations which, in addition, satisfy four constraints which are purely algebraic in field strength; one has an imbedding of the Einstein constraint surface into the Yang-Mills theory. This imbedding preserves the Poisson-bracket structure of the two theories. On the other hand, it does not commute with time evolution; the Yang-Mills Hamiltonian is very different from Einstein's. However, since the Einstein Hamiltonian is a linear combination of constraints and a surface term, the simplification of constraints is significant also for Einstein dynamics.

To go over to quantum theory, as in the Yang-Mills case, we shall replace the basic variables  $\pm A_a$  and  $\tilde{\sigma}^a \equiv E^a$  by operators  $\pm \hat{A}_a$  and  $\hat{\sigma}^a$  satisfying the canonical commutation relation

$$\begin{aligned} &\{ \pm \hat{A}_m^{MN}(x), \hat{\sigma}^a_{AB}(y) \} \\ &= (\hbar/\sqrt{2}) \delta_m^a \delta_A^M \delta_B^N \delta(x, y). \quad (12) \end{aligned}$$

The shift of variables from  $(\hat{q}, \hat{p})$  to  $(\hat{\sigma}, \pm \hat{A})$  simplifies several issues in the quantum theory as a result of the features of constraints (4')–(6'), noted below (5'). I now summarize the construction and the results that follow.

First, one can ask if there exists a factor ordering for the quantum version of constraints (4')–(6') for which the quantum constraints are closed under the commutator bracket, i.e., for which the evaluation of commutators yields a result in which a constraint operator always appears on the right. The answer is in the affirmative. Set

$$\hat{C}_N := (\sqrt{2}/\hbar) \int_{\Sigma} G^{-1} N_A{}^B (\pm \hat{D}_a \hat{\sigma}^a)_B{}^A + N^a \text{Tr} \hat{\sigma}^b \pm \hat{F}_{ab} + N \text{Tr} \hat{\sigma}^a \hat{\sigma}^b \pm \hat{F}_{ab}, \quad (13)$$

where  $N$  stands for the triplet of smearing fields  $(N^a{}_B, N^a, N)$ . By use of (12) it then follows that

$$[\hat{C}_N, \hat{C}_M] = \hat{C}_P,$$

where

$$\begin{aligned} P_A{}^B &= [N, M]_A{}^B + G^{-1} N^a M^b \pm \hat{F}_{ab}{}_A{}^B + G^{-1} (MN^a - NM^a) (\hat{\sigma}^b{}_A{}^M \pm \hat{F}_{ab}{}_M{}^B - \hat{\sigma}^b{}_M{}^B \pm \hat{F}_{ab}{}_A{}^M), \\ P^a &= \mathcal{L}_M N^a + 2(N D_b M - M D_b N) \hat{q}^{ab}, \quad P = \mathcal{L}_M N - \mathcal{L}_N M. \end{aligned} \quad (14)$$

In this result, the presence of internal indices and the consequent constraint (6') plays a crucial role: Certain unwanted terms vanish because of the symmetry properties of their internal indices, and the commutators of (4') and (5') involve not only these constraints but also (6'). Also, the presence of the internal indices in (4')–(6') (as well as of  $\hat{\sigma}^a$ , the “square root” of  $\hat{q}^{ab}$ ) makes it impossible to translate the preferred factor ordering in terms of the traditional variables  $(\hat{q}_{ab}, \hat{p}^{ab})$ . Note, however, that the argument is only *formal*; I have not regularized the products of operator-valued distributions in (4')–(6'). Nonetheless, formal closure is significant since it can fail even for systems with a finite number of degrees of freedom<sup>6</sup> where the issues of regularization never arise.

Next, we come to the issue of finding representations of the canonical commutation relation (12). Since, as noted above, the constraints are at worst quadratic in each of  $\hat{\sigma}^a$  and  $\hat{A}_a$ , it is feasible to study both the  $\hat{\sigma}$  representation, in which quantum states are general complex-valued functionals of  $\hat{\sigma}$ , and the  $\pm A_a$  representation, in which they are holomorphic functionals of  $\pm A_a$ . [The analogs for a simple harmonic oscillator are respectively the position representation,  $\psi \equiv \psi(x)$ , and the Bargmann representation,  $\psi \equiv \psi(z)$ ,  $z = x + ip$ .] This is in striking contrast to the situation with  $(q_{ab}, p^{ab})$  variables, where the momentum representation is unmanageable because the constraints have a complicated  $q$  dependence.

Let us focus on the  $\pm A_a$  representation since it enables one to borrow some ideas from (quantum) Yang-Mills theory. The weak-field limit has been studied in detail. Here, the quantum constraints are solved precisely by the requirement that the states be holomorphic functionals of the symmetric, trace-free, transverse part  $(\delta A_a^i)^{STT}$  of the linearized connection  $(\delta^\pm A_a)$ , and the Hamiltonian is given simply by

$$H = (G/4\pi) \int_{\Sigma} (\delta A)_{ab}^{STT} (\delta A)_{STT}^{tab} d^3x \quad (15)$$

(which is analogous to the expression  $H = ZZ^*$  for the

simple harmonic oscillator). In exact theory the representation is being investigated by Jacobson and Smolin.<sup>7</sup>

Next, note that the constraints (4'')–(6'') in terms of the Yang-Mills variables  $(\pm A_a, E^a)$  do not require a background structure such as a metric or a derivative operator. (This is to be contrasted with the Yang-Mills evolution equations which do require a background metric.) Consequently, one can take over techniques from the Hamiltonian lattice QCD<sup>8</sup> to put the quantum gravity on a lattice. The advantage of a lattice formulation is that the constraints do not have to be regularized. This line of investigation is being pursued by Renteln and Smolin.<sup>9</sup>

Finally, I have restricted the discussion to the vacuum case only for simplicity. It is straightforward to include a cosmological constant and matter sources—Yang-Mills sources fit in especially well—in the framework.

Details will appear elsewhere.

I am most grateful to Ted Jacobson, Lee Smolin, and Paul Renteln for discussions. This work was supported by National Science Foundation Grants No. PHY-83-10041 and No. PHY82-17853, supplemented by funds from the U. S. National Aeronautics and Space Administration.

<sup>1</sup>See, e.g., K. Kuchař, in *Quantum Gravity 2*, edited by C. J. Isham, R. Penrose, and D. W. Sciama (Oxford Univ. Press, New York, 1981).

<sup>2</sup>Note that, when (6) holds,  $\Pi_{ab}$  is just the extrinsic curvature.

<sup>3</sup>A. Ashtekar, T. Jacobson, and L. Smolin, to be published.

<sup>4</sup>See, e.g., M. Ko, M. Ludvigsen, E. T. Newman, and K. P. Tod, Phys. Rep. **71**, 51 (1981).

<sup>5</sup>R. S. Ward, “Integrable and Solvable Systems” (to be published).

<sup>6</sup>See, e.g., A. Ashtekar and M. Stillerman, J. Math. Phys.

(N.Y.) **27**, 1319 (1986).

<sup>7</sup>They construct quantum states from the holonomy operator of  $\pm A_a$  associated with closed loops and regularize the action of constraints on these loop states. Their results indicate that the metric  $q^{ab}$  would be degenerate microscopically and acquire its usual geometrical meaning only on macroscopic regions and in

states with a large number of loops. (T. Jacobson and L. Smolin, private communication.)

<sup>8</sup>See, e.g., L. Susskind, in *Weak and Electromagnetic Interactions*, edited by R. Balian and C. H. Llewellyn Smith (North-Holland, Amsterdam, 1977).

<sup>9</sup>P. Renteln and L. Smolin, private communication.

## Teleporting an Unknown Quantum State via Dual Classical and Einstein-Podolsky-Rosen Channels

Charles H. Bennett,<sup>(1)</sup> Gilles Brassard,<sup>(2)</sup> Claude Crépeau,<sup>(2),(3)</sup>

Richard Jozsa,<sup>(2)</sup> Asher Peres,<sup>(4)</sup> and William K. Wootters<sup>(5)</sup>

<sup>(1)</sup>*IBM Research Division, T.J. Watson Research Center, Yorktown Heights, New York 10598*

<sup>(2)</sup>*Département IRO, Université de Montréal, C.P. 6128, Succursale "A", Montréal, Québec, Canada H3C 3J7*

<sup>(3)</sup>*Laboratoire d'Informatique de l'École Normale Supérieure, 45 rue d'Ulm, 75230 Paris CEDEX 05, France<sup>(a)</sup>*

<sup>(4)</sup>*Department of Physics, Technion-Israel Institute of Technology, 32000 Haifa, Israel*

<sup>(5)</sup>*Department of Physics, Williams College, Williamstown, Massachusetts 01267*

(Received 2 December 1992)

An unknown quantum state  $|\phi\rangle$  can be disassembled into, then later reconstructed from, purely classical information and purely nonclassical Einstein-Podolsky-Rosen (EPR) correlations. To do so the sender, "Alice," and the receiver, "Bob," must prearrange the sharing of an EPR-correlated pair of particles. Alice makes a joint measurement on her EPR particle and the unknown quantum system, and sends Bob the classical result of this measurement. Knowing this, Bob can convert the state of his EPR particle into an exact replica of the unknown state  $|\phi\rangle$  which Alice destroyed.

PACS numbers: 03.65.Bz, 42.50.Dv, 89.70.+c

The existence of long range correlations between Einstein-Podolsky-Rosen (EPR) [1] pairs of particles raises the question of their use for information transfer. Einstein himself used the word "telepathically" in this context [2]. It is known that *instantaneous* information transfer is definitely impossible [3]. Here, we show that EPR correlations can nevertheless assist in the "teleportation" of an intact quantum state from one place to another, by a sender who knows neither the state to be teleported nor the location of the intended receiver.

Suppose one observer, whom we shall call "Alice," has been given a quantum system such as a photon or spin- $\frac{1}{2}$  particle, prepared in a state  $|\phi\rangle$  unknown to her, and she wishes to communicate to another observer, "Bob," sufficient information about the quantum system for him to make an accurate copy of it. Knowing the state vector  $|\phi\rangle$  itself would be sufficient information, but in general there is no way to learn it. Only if Alice knows beforehand that  $|\phi\rangle$  belongs to a given orthonormal set can she make a measurement whose result will allow her to make an accurate copy of  $|\phi\rangle$ . Conversely, if the possibilities for  $|\phi\rangle$  include two or more nonorthogonal states, then no measurement will yield sufficient information to prepare

a perfectly accurate copy.

A trivial way for Alice to provide Bob with all the information in  $|\phi\rangle$  would be to send the particle itself. If she wants to avoid transferring the original particle, she can make it interact unitarily with another system, or "ancilla," initially in a known state  $|a_0\rangle$ , in such a way that after the interaction the original particle is left in a standard state  $|\phi_0\rangle$  and the ancilla is in an unknown state  $|a\rangle$  containing complete information about  $|\phi\rangle$ . If Alice now sends Bob the ancilla (perhaps technically easier than sending the original particle), Bob can reverse her actions to prepare a replica of her original state  $|\phi\rangle$ . This "spin-exchange measurement" [4] illustrates an essential feature of quantum information: it can be swapped from one system to another, but it cannot be duplicated or "cloned" [5]. In this regard it is quite unlike classical information, which can be duplicated at will. The most tangible manifestation of the nonclassicality of quantum information is the violation of Bell's inequalities [6] observed [7] in experiments on EPR states. Other manifestations include the possibility of quantum cryptography [8], quantum parallel computation [9], and the superiority of interactive measurements for extracting informa-

tion from a pair of identically prepared particles [10].

The spin-exchange method of sending full information to Bob still lumps classical and nonclassical information together in a single transmission. Below, we show how Alice can divide the full information encoded in  $|\phi\rangle$  into two parts, one purely classical and the other purely nonclassical, and send them to Bob through two different channels. Having received these two transmissions, Bob can construct an accurate replica of  $|\phi\rangle$ . Of course Alice's original  $|\phi\rangle$  is destroyed in the process, as it must be to obey the no-cloning theorem. We call the process we are about to describe teleportation, a term from science fiction meaning to make a person or object disappear while an exact replica appears somewhere else. It must be emphasized that our teleportation, unlike some science fiction versions, defies no physical laws. In particular, it cannot take place instantaneously or over a spacelike interval, because it requires, among other things, sending a classical message from Alice to Bob. The net result of teleportation is completely prosaic: the removal of  $|\phi\rangle$  from Alice's hands and its appearance in Bob's hands a suitable time later. The only remarkable feature is that, in the interim, the information in  $|\phi\rangle$  has been cleanly separated into classical and nonclassical parts. First we shall show how to teleport the quantum state  $|\phi\rangle$  of a spin- $\frac{1}{2}$  particle. Later we discuss teleportation of more complicated states.

The nonclassical part is transmitted first. To do so, two spin- $\frac{1}{2}$  particles are prepared in an EPR singlet state

$$|\Psi_{23}^{(-)}\rangle = \sqrt{\frac{1}{2}} (|\uparrow_2\rangle|\downarrow_3\rangle - |\downarrow_2\rangle|\uparrow_3\rangle). \quad (1)$$

The subscripts 2 and 3 label the particles in this EPR pair. Alice's original particle, whose unknown state  $|\phi\rangle$  she seeks to teleport to Bob, will be designated by a subscript 1 when necessary. These three particles may be of different kinds, e.g., one or more may be photons, the polarization degree of freedom having the same algebra as a spin.

One EPR particle (particle 2) is given to Alice, while

$$|\Psi_{123}\rangle = \frac{1}{2} [|\Psi_{12}^{(-)}\rangle (-a|\uparrow_3\rangle - b|\downarrow_3\rangle) + |\Psi_{12}^{(+)}\rangle (-a|\uparrow_3\rangle + b|\downarrow_3\rangle) + |\Phi_{12}^{(-)}\rangle (a|\downarrow_3\rangle + b|\uparrow_3\rangle) + |\Phi_{12}^{(+)}\rangle (a|\downarrow_3\rangle - b|\uparrow_3\rangle)]. \quad (5)$$

It follows that, regardless of the unknown state  $|\phi_1\rangle$ , the four measurement outcomes are equally likely, each occurring with probability 1/4. Furthermore, after Alice's measurement, Bob's particle 3 will have been projected into one of the four pure states superposed in Eq. (5), according to the measurement outcome. These are, respectively,

$$\begin{aligned} -|\phi_3\rangle &\equiv -\begin{pmatrix} a \\ b \end{pmatrix}, \quad \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}|\phi_3\rangle, \\ &\quad \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}|\phi_3\rangle, \quad \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}|\phi_3\rangle. \end{aligned} \quad (6)$$

the other (particle 3) is given to Bob. Although this establishes the possibility of nonclassical correlations between Alice and Bob, the EPR pair at this stage contains no information about  $|\phi\rangle$ . Indeed the entire system, comprising Alice's unknown particle 1 and the EPR pair, is in a pure product state,  $|\phi_1\rangle|\Psi_{23}^{(-)}\rangle$ , involving neither classical correlation nor quantum entanglement between the unknown particle and the EPR pair. Therefore no measurement on either member of the EPR pair, or both together, can yield any information about  $|\phi\rangle$ . An entanglement between these two subsystems is brought about in the next step.

To couple the first particle with the EPR pair, Alice performs a complete measurement of the von Neumann type on the joint system consisting of particle 1 and particle 2 (her EPR particle). This measurement is performed in the Bell operator basis [11] consisting of  $|\Psi_{12}^{(-)}\rangle$  and

$$\begin{aligned} |\Psi_{12}^{(+)}\rangle &= \sqrt{\frac{1}{2}} (|\uparrow_1\rangle|\downarrow_2\rangle + |\downarrow_1\rangle|\uparrow_2\rangle), \\ |\Phi_{12}^{(\pm)}\rangle &= \sqrt{\frac{1}{2}} (|\uparrow_1\rangle|\uparrow_2\rangle \pm |\downarrow_1\rangle|\downarrow_2\rangle). \end{aligned} \quad (2)$$

Note that these four states are a complete orthonormal basis for particles 1 and 2.

It is convenient to write the unknown state of the first particle as

$$|\phi_1\rangle = a|\uparrow_1\rangle + b|\downarrow_1\rangle, \quad (3)$$

with  $|a|^2 + |b|^2 = 1$ . The complete state of the three particles before Alice's measurement is thus

$$\begin{aligned} |\Psi_{123}\rangle &= \frac{a}{\sqrt{2}} (|\uparrow_1\rangle|\uparrow_2\rangle|\downarrow_3\rangle - |\uparrow_1\rangle|\downarrow_2\rangle|\uparrow_3\rangle) \\ &\quad + \frac{b}{\sqrt{2}} (|\downarrow_1\rangle|\uparrow_2\rangle|\downarrow_3\rangle - |\downarrow_1\rangle|\downarrow_2\rangle|\uparrow_3\rangle). \end{aligned} \quad (4)$$

In this equation, each direct product  $|\downarrow_1\rangle|\downarrow_2\rangle$  can be expressed in terms of the Bell operator basis vectors  $|\Phi_{12}^{(\pm)}\rangle$  and  $|\Psi_{12}^{(\pm)}\rangle$ , and we obtain

Each of these possible resultant states for Bob's EPR particle is related in a simple way to the original state  $|\phi\rangle$  which Alice sought to teleport. In the case of the first (singlet) outcome, Bob's state is the same except for an irrelevant phase factor, so Bob need do nothing further to produce a replica of Alice's spin. In the three other cases, Bob must apply one of the unitary operators in Eq. (6), corresponding, respectively, to 180° rotations around the  $z$ ,  $x$ , and  $y$  axes, in order to convert his EPR particle into a replica of Alice's original state  $|\phi\rangle$ . (If  $|\phi\rangle$  represents a photon polarization state, a suitable combination of half-

wave plates will perform these unitary operations.) Thus an accurate teleportation can be achieved in all cases by having Alice tell Bob the classical outcome of her measurement, after which Bob applies the required rotation to transform the state of his particle into a replica of  $|\phi\rangle$ . Alice, on the other hand, is left with particles 1 and 2 in one of the states  $|\Psi_{12}^{(\pm)}\rangle$  or  $|\Phi_{12}^{(\pm)}\rangle$ , without any trace of the original state  $|\phi\rangle$ .

Unlike the quantum correlation of Bob's EPR particle 3 to Alice's particle 2, the result of Alice's measurement is purely classical information, which can be transmitted, copied, and stored at will in any suitable physical medium. In particular, this information need not be destroyed or canceled to bring the teleportation process to a successful conclusion: The teleportation of  $|\phi\rangle$  from Alice to Bob has the side effect of producing two bits of random classical information, uncorrelated to  $|\phi\rangle$ , which are left behind at the end of the process.

Since teleportation is a *linear* operation applied to the quantum state  $|\phi\rangle$ , it will work not only with pure states, but also with mixed or entangled states. For example, let Alice's original particle 1 be itself part of an EPR singlet with another particle, labeled 0, which may be far away from both Alice and Bob. Then, after teleportation, particles 0 and 3 would be left in a singlet state, even though they had originally belonged to separate EPR pairs.

All of what we have said above can be generalized to systems having  $N > 2$  orthogonal states. In place of an EPR spin pair in the singlet state, Alice would use a pair of  $N$ -state particles in a completely entangled state. For definiteness let us write this entangled state as  $\sum_j |j\rangle \otimes |j\rangle / \sqrt{N}$ , where  $j = 0, 1, \dots, N-1$  labels the  $N$  elements of an orthonormal basis for each of the  $N$ -state systems. As before, Alice performs a joint measurement on particles 1 and 2. One such measurement that has the desired effect is the one whose eigenstates are  $|\psi_{nm}\rangle$ , defined by

$$|\psi_{nm}\rangle = \sum_j e^{2\pi i j n / N} |j\rangle \otimes |(j+m) \bmod N\rangle / \sqrt{N}. \quad (7)$$

Once Bob learns from Alice that she has obtained the result  $nm$ , he performs on his previously entangled particle (particle 3) the unitary transformation

$$U_{nm} = \sum_k e^{2\pi i k n / N} |k\rangle \langle (k+m) \bmod N|. \quad (8)$$

This transformation brings Bob's particle to the original state of Alice's particle 1, and the teleportation is complete.

The classical message plays an essential role in teleportation. To see why, suppose that Bob is impatient, and tries to complete the teleportation by guessing Alice's classical message before it arrives. Then Alice's expected

$|\phi\rangle$  will be reconstructed (in the spin- $\frac{1}{2}$  case) as a random mixture of the four states of Eq. (6). For any  $|\phi\rangle$ , this is a maximally mixed state, giving no information about the input state  $|\phi\rangle$ . It could not be otherwise, because any correlation between the input and the guessed output could be used to send a superluminal signal.

One may still inquire whether accurate teleportation of a two-state particle requires a full two bits of classical information. Could it be done, for example, using only two or three distinct classical messages instead of four, or four messages of unequal probability? Later we show that a full two bits of classical channel capacity are necessary. Accurate teleportation using a classical channel of any lesser capacity would allow Bob to send superluminal messages through the teleported particle, by guessing the classical message before it arrived (cf. Fig. 2).

Conversely one may inquire whether other states besides an EPR singlet can be used as the nonclassical channel of the teleportation process. Clearly any direct product state of particles 2 and 3 is useless, because for such states manipulation of particle 2 has no effect on what can be predicted about particle 3. Consider now a non-factorable state  $|\Upsilon_{23}\rangle$ . It can readily be seen that after Alice's measurement, Bob's particle 3 will be related to  $|\phi_1\rangle$  by four fixed unitary operations if and only if  $|\Upsilon_{23}\rangle$  has the form

$$\sqrt{\frac{1}{2}}(|u_2\rangle|p_3\rangle + |v_2\rangle|q_3\rangle), \quad (9)$$

where  $\{|u\rangle, |v\rangle\}$  and  $\{|p\rangle, |q\rangle\}$  are any two pairs of orthonormal states. These are maximally entangled states [11], having maximally random marginal statistics for measurements on either particle separately. States which are less entangled reduce the fidelity of teleportation, and/or the range of states  $|\phi\rangle$  that can be accurately teleported. The states in Eq. (9) are also precisely those obtainable from the EPR singlet by a local one-particle unitary operation [12]. Their use for the nonclassical channel is entirely equivalent to that of the singlet (1). Maximal entanglement is necessary and sufficient for faithful tele-

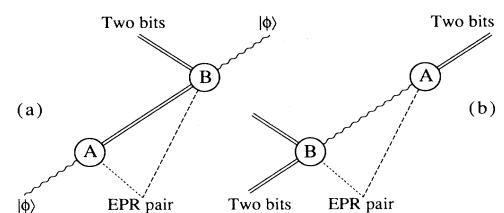


FIG. 1. Spacetime diagrams for (a) quantum teleportation, and (b) 4-way coding [12]. As usual, time increases from bottom to top. The solid lines represent a classical pair of bits, the dashed lines an EPR pair of particles (which may be of different types), and the wavy line a quantum particle in an unknown state  $|\phi\rangle$ . Alice (A) performs a quantum measurement, and Bob (B) a unitary operation.

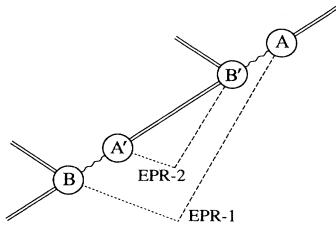


FIG. 2. Spacetime diagram of a more complex 4-way coding scheme in which the modulated EPR particle (wavy line) is teleported rather than being transmitted directly. This diagram can be used to prove that a classical channel of two bits of capacity is necessary for teleportation. To do so, assume on the contrary that the teleportation from  $A'$  to  $B'$  uses an internal classical channel of capacity  $C < 2$  bits, but is still able to transmit the wavy particle's state accurately from  $A'$  to  $B'$ , and therefore still transmit the external two-bit message accurately from  $B$  to  $A$ . The assumed lower capacity  $C < 2$  of the internal channel means that if  $B'$  were to guess the internal classical message superluminally instead of waiting for it to arrive, his probability  $2^{-C}$  of guessing correctly would exceed  $1/4$ , resulting in a probability greater than  $1/4$  for successful superluminal transmission of the external two-bit message from  $B$  to  $A$ . This in turn entails the existence of two distinct external two-bit messages,  $r$  and  $s$ , such that  $P(r|s)$ , the probability of superluminally receiving  $r$  if  $s$  was sent, is less than  $1/4$ , while  $P(r|r)$ , the probability of superluminally receiving  $r$  if  $r$  was sent, is greater than  $1/4$ . By redundant coding, even this statistical difference between  $r$  and  $s$  could be used to send reliable superluminal messages; therefore reliable teleportation of a two-state particle cannot be achieved with a classical channel of less than two bits of capacity. By the same argument, reliable teleportation of an  $N$ -state particle requires a classical channel of  $2\log_2(N)$  bits capacity.

portation.

Although it is currently unfeasible to store separated EPR particles for more than a brief time, if it becomes feasible to do so, quantum teleportation could be quite useful. Alice and Bob would only need a stockpile of EPR pairs (whose reliability can be tested by violations of Bell's inequality [7]) and a channel capable of carrying robust classical messages. Alice could then teleport quantum states to Bob over arbitrarily great distances, without worrying about the effects of attenuation and noise on, say, a single photon sent through a long optical fiber. As an application of teleportation, consider the problem investigated by Peres and Wootters [10], in which Bob already has another copy of  $|\phi\rangle$ . If he acquires Alice's copy, he can measure both together, thereby determining the state  $|\phi\rangle$  more accurately than can be done by making a separate measurement on each one. Finally, teleportation has the advantage of still being possible in situations where Alice and Bob, after sharing their EPR pairs, have wandered about independently and no longer know each others' locations. Alice cannot reliably send

Bob the original quantum particle, or a spin-exchanged version of it, if she does not know where he is; but she can still teleport the quantum state to him, by broadcasting the classical information to all places where he might be.

Teleportation resembles another recent scheme for using EPR correlations to help transmit useful information. In "4-way coding" [12] modulation of one member of an EPR pair serves to reliably encode a 2-bit message in the joint state of the complete pair. Teleportation and 4-way coding can be seen as variations on the same underlying process, illustrated by the spacetime diagrams in Fig. 1. Note that *closed loops* are involved for both processes. Trying to draw similar "Feynman diagrams" with tree structure, rather than loops, would lead to physically impossible processes.

On the other hand, more complicated closed-loop diagrams are possible, such as Fig. 2, obtained by substituting Fig. 1(a) into the wavy line of Fig. 1(b). This represents a 4-way coding scheme in which the modulated EPR particle is teleported instead of being transmitted directly. Two incoming classical bits on the lower left are reproduced reliably on the upper right, with the assistance of two shared EPR pairs and two other classical bits, uncorrelated with the external bits, in an internal channel from  $A'$  to  $B'$ . This diagram is of interest because it can be used to show that a full two bits of classical channel capacity are necessary for accurate teleportation of a two-state particle (cf. caption).

Work by G.B. is supported by NSERC's E. W. R. Steacie Memorial Fellowship and Québec's FCAR. A.P. was supported by the Gerard Swope Fund and the Fund for Encouragement of Research. Laboratoire d'Informatique de l'Ecole Normale Supérieure is associée au CNRS URA 1327.

(a) Permanent address.

- [1] A. Einstein, B. Podolsky, and N. Rosen, Phys. Rev. **47**, 777 (1935); D. Bohm, *Quantum Theory* (Prentice Hall, Englewood Cliffs, NJ, 1951).
- [2] A. Einstein, in *Albert Einstein, Philosopher-Scientist*, edited by P. A. Schilpp (Library of Living Philosophers, Evanston, 1949) p. 85.
- [3] A. Shimony, in *Proceedings of the International Symposium on Foundations of Quantum Theory* (Physical Society of Japan, Tokyo, 1984).
- [4] J. L. Park, Found. Phys. **1**, 23 (1970).
- [5] W. K. Wootters and W. H. Zurek, Nature (London) **299**, 802 (1982).
- [6] J. S. Bell, Physics (Long Island City, N.Y.) **1**, 195 (1964); J. F. Clauser, M. A. Horne, A. Shimony, and R. A. Holt, Phys. Rev. Lett. **23**, 880 (1969).
- [7] A. Aspect, J. Dalibard, and G. Roger, Phys. Rev. Lett. **49**, 1804 (1982); Y. H. Shih and C. O. Alley, Phys. Rev. Lett. **61**, 2921 (1988).
- [8] S. Wiesner, Sigact News **15**, 78 (1983); C. H. Bennett and G. Brassard, in *Proceedings of IEEE International*

- Conference on Computers, Systems, and Signal Processing, Bangalore, India* (IEEE, New York, 1984), pp. 175–179; A. K. Ekert, Phys. Rev. Lett. **67**, 661 (1991); C. H. Bennett, G. Brassard, and N. D. Mermin, Phys. Rev. Lett. **68**, 557–559 (1992); C. H. Bennett, Phys. Rev. Lett. **68**, 3121 (1992); A. K. Ekert, J. G. Rarity, P. R. Tapster, and G. M. Palma, Phys. Rev. Lett. **69**, 1293 (1992); C. H. Bennett, G. Brassard, C. Crépeau, and M.-H. Skubiszewska, *Advances in Cryptology—Crypto '91 Proceedings, August 1991* (Springer, New York, 1992), pp. 351–366; G. Brassard and C. Crépeau, *Advances in Cryptology—Crypto '90 Proceedings, August 1990* (Springer, New York, 1991), pp. 49–61.
- [9] D. Deutsch Proc. R. Soc. London A **400**, 97 (1985);  
D. Deutsch and R. Jozsa, Proc. R. Soc. London A **439**, 553–558 (1992); A. Berthiaume and G. Brassard, in *Proceedings of the Seventh Annual IEEE Conference on Structure in Complexity Theory, Boston, June 1992*, (IEEE, New York, 1989), pp. 132–137; “Oracle Quantum Computing,” Proceedings of the Workshop on Physics and Computation, PhysComp 92, (IEEE, Dallas, to be published).
- [10] A. Peres and W. K. Wootters, Phys. Rev. Lett. **66**, 1119 (1991).
- [11] S. L. Braunstein, A. Mann, and M. Revzen, Phys. Rev. Lett. **68**, 3259 (1992).
- [12] C. H. Bennett and S. J. Wiesner, Phys. Rev. Lett. **69**, 2881 (1992).



THU-93/26  
gr-qc/9310026

## DIMENSIONAL REDUCTION in QUANTUM GRAVITY<sup>†</sup>

G. 't Hooft

Institute for Theoretical Physics

Utrecht University

Postbox 80 006, 3508 TA Utrecht, the Netherlands

### Abstract

The requirement that physical phenomena associated with gravitational collapse should be duly reconciled with the postulates of quantum mechanics implies that at a Planckian scale our world is not 3+1 dimensional. Rather, the observable degrees of freedom can best be described as if they were Boolean variables defined on a two-dimensional lattice, evolving with time. This observation, deduced from not much more than unitarity, entropy and counting arguments, implies severe restrictions on possible models of quantum gravity. Using cellular automata as an example it is argued that this dimensional reduction implies more constraints than the freedom we have in constructing models. This is the main reason why so-far no completely consistent mathematical models of quantum black holes have been found.

---

<sup>†</sup> Essay dedicated to Abdus Salam

With the request to write a short paper in honor of Abdus Salam I am given the opportunity to contemplate some very deep questions concerning the ultimate unification that may perhaps be achieved when all aspects of quantum theory, particle theory and general relativity are combined. One of these questions is the dimensionality of space and time.

At first sight our world has three spacelike dimensions and one timelike. This was used as a starting point of all quantum field theories and indeed also all string theories as soon as they invoke the Kaluza Klein mechanism. Also when we quantize gravity perturbatively we start by postulating a Fock space in which basically free particles roam in a three plus one dimensional world. Naturally, when people discuss possible cut-off mechanisms, they think of some sort of lattice scheme either in 3+1 dimensional Minkowski space or in 4 dimensional Euclidean space. The cut-off distance scale is then suspected to be the Planck scale.

Unfortunately any such lattice scheme seems to be in conflict with local Lorentz invariance or Euclidean invariance, as the case may be, and most of all also with coordinate reparametrization invariance. It seems to be virtually impossible to recover these symmetries at large distance scales, where we want them. So the details of the cut-off are kept necessarily vague.

The most direct and obvious *physical* cut-off does not come from non-renormalizability alone, but from the formation of microscopic black holes as soon as too much energy would be accumulated into too small a region. From a physical point of view it is the black holes that should provide for a natural cut-off all by themselves.

This has been this author's main subject of research for over a decade. A mathematically consistent formulation of the black hole cut-off turns out to be extremely difficult to find, and in this short note I will explain what may well be the main reason for this difficulty: nature is much more crazy at the Planck scale than even string theorists could have imagined.

One of my starting points has been that quantum mechanics itself is not at all a mystery to me. The emergence of a Hilbert space with a Copenhagen interpretation of its inner products is a quite natural feature of any theory with the following characteristics at a local scale: the system must have *discrete* degrees of freedom at tiny distance scales, and the laws of evolution must be *reversible* in time. With discrete degrees of freedom one can construct Hilbert space in a quite natural way by postulating that any state of the physical degrees of freedom corresponds to an element of a basis of this Hilbert space[1]. Reversibility in time is required if we wish to see a quantum superposition principle: the norm of all states is then preserved, even if they are quantum superpositions of these basis elements.

Needless to say, one might suspect that some or all of the quantum mechanical postulates could break down at the Planck scale[2]. But then one might as well throw away anything we know about physics, and that is not the route I want to follow. I have never seen convincing models where ordinary quantum mechanics breaks down at a microscopic level but is somehow recovered at the atomic scale. Therefore I prefer not to speculate that quantum mechanics breaks down at the Planck scale, but instead to suspect that quantum mechanics becomes trivial there: quantum superpositions are still allowed there but become irrelevant.

Black holes then present a major challenge. At first sight they render time reversibility impossible. Objects thrown into a black hole can never be retrieved[3]. Things are often presented as if black holes connect our world to other universes via wormholes[4], or, if one prefers not to refer to these other universes one says that information thrown in can not be retrieved anymore[5]. According to this view information is preserved only if one considers multiply connected universes[6]. This is useless if one wishes to recover quantum mechanical behavior in our universe by itself. However, at closer inspection one finds that any object thrown into a black hole actually does leave some signals behind in our own world[7], and it is conceivable that a unitary theory for our own universe can be built using this as a starting point.[8]

The main difficulty then is to formulate exactly what our degrees of freedom are. Remarkably, it is relatively easy to give a fairly precise estimate of *how many* degrees of freedom we have. This can be deduced in two equivalent ways[9, 10]. One is by considering capture and emission of objects by black holes as scattering experiments. If these are described quantum mechanically one can deduce information about the size of phase space. It must be finite, increasing exponentially with the surface area of the black hole horizon. The other way to deduce the same information is by using thermodynamics. One derives the entropy  $S$  of a black hole, finding

$$S = 4\pi M^2 + C, \quad (1)$$

in natural units. The constant  $C$  is not known (in fact there could be as yet unknown subdominant terms in this expression, increasing slower than  $M^2$  as the mass increases). In principle  $C$  could be infinite (even for small  $M$ ), which basically corresponds to the remnant theory. I think an infinite  $C$  would induce major problems (again implying a deviation from ordinary quantum mechanical behavior) in a consistent theory, so I will henceforth assume  $C$  to be bounded.

This entropy is often attributed ‘to the space-time metric itself’, as if there would be yet another separate contribution from the quantized fields in this metric, such as the contribution to the entropy of objects outside the black hole. However, this would not be correct. First of all, the contribution of objects at some distance from the black hole will

always be negligible unless they are really very far away (at  $|x| \gg R_{\text{Schwarzschild}}$ ). But most importantly, the contribution of quantum fields to the entropy *diverges* very near the horizon[10]. The reason for this divergence can be easily understood physically: arbitrary amounts of matter can be thrown in and arbitrary amounts are waiting to radiate out; they contribute infinitely to the total number of physical degrees of freedom.

Yet the black hole entropy had just been argued to be finite, *even with the quantum fields present!* We conclude that one has to attribute the black hole entropy not to the space-time metric itself but to the quantized fields present there (which of course does include the small quantum fluctuations of the metric itself), and then one must choose a cut-off sufficiently close to the horizon such that it exactly matches the known black hole entropy (1). In short, the black hole entropy *includes* the entropy of the quantized fields in its neighborhood[10].

In any quantum theory there is a ‘third law of thermodynamics’ relating the entropy to the total number of degrees of freedom: the dimension of the vector space describing all possible states our system can be in is the exponent of the entropy. For instance in a discrete theory described by  $n$  spins that can take only two values (‘Boolean variables’), the dimension  $\mathcal{N}$  of Hilbert space is

$$e^S = \mathcal{N} = 2^n , \quad (2)$$

Hence the entropy directly counts the number of Boolean degrees of freedom.

Considering the fact that at the Hawking temperature the contribution to the entropy of fields anywhere in the region  $R < |x| \ll R^3$  pales compared to the black hole entropy itself, we can now make an important observation concerning the relevant degrees of freedom of a black hole:

*The total number of Boolean degrees of freedom,  $n$ , in a region of space-time surrounding a black hole is*

$$n = \frac{S}{\ln 2} = \frac{4\pi M^2}{\ln 2} = \frac{A}{4 \ln 2} , \quad (3)$$

*where  $A$  is the horizon area.*

We can carry this argument one step further. Consider just any closed spacelike surface, with  $S(2)$  topology and total surface area  $A$ . Consider all possible field and metric configurations inside this surface. We ask how many mutually orthogonal states there can be. If we want these states to be observable for the outside world we have to insist that the total energy inside the surface be less than 1/4 times its linear dimensions, otherwise our surface would lie within the Schwarzschild radius. Let us first ask how many states would an ordinary quantum field theory allow us to have, given these limits on the volume  $V$  and the energy  $E$ . Now this is not hard. The most probable state would be a

gas at some temperature  $T = 1/\beta$ . Its energy would be approximately

$$E = C_1 Z V T^4, \quad (4)$$

where  $Z$  is the number of different fundamental particle types with mass less than  $T$  and  $C_1$  a numerical constant of order one, all in natural units. The total entropy  $S$  is

$$S = C_2 Z V T^3, \quad (5)$$

where  $C_2$  is another dimensionless constant. Now the Schwarzschild limit requires that

$$2E < (V/(\frac{4}{3}\pi))^{\frac{1}{3}}, \quad (6)$$

hence, with eq. (4),

$$T < C_3 Z^{-\frac{1}{4}} V^{-\frac{1}{6}}, \quad (7)$$

so that

$$S < C_4 Z^{\frac{1}{4}} V^{\frac{1}{2}} = C_5 Z^{\frac{1}{4}} A^{\frac{3}{4}}. \quad (8)$$

The  $C_i$  are all constants of order 1 in natural units. Since in quantum field theories, at sufficiently low temperatures,  $Z$  is limited by a dimensionless number we find that this entropy is small compared to that of a black hole, if the area  $A$  is sufficiently large.

Next consider a set of  $N$  black holes, with masses  $M_i$ . They contribute to the energy  $\sum_i M_i$ . So

$$\sum_i M_i < C_6 A^{\frac{1}{2}}, \quad (9)$$

while their total entropy is given by

$$S = C_7 \sum_i M_i^2, \quad (10)$$

where we note that the contribution of their movements to the entropy is negligible. We see that ineq. (10) is saturated when one black hole has the largest possible size that still fits inside our area. Its entropy is

$$S_{max} = \frac{1}{4} A, \quad (11)$$

and this is as large as we can ever make it. This therefore answers our question concerning the total number of possible states. It is given by eqs. (2) and (3). The single black hole is the limit[11].

The importance of this result can hardly be overestimated. At first sight it is counterintuitive. One would have expected that the number of possible states would grow exponentially with the volume, not the area, as in any ordinary field theory with a cut-off.

So if one would take a regularized fermion theory with cut-off at the Planck scale one would get far too many states. But it is clear why our answer came out this way. Most of the states of a regularized quantum field theory would have so much energy that they would collapse into a black hole before they could dictate the further evolution of the system in time. If we want to avoid this a much more rigorous cut-off than a Planckian one must be called for. Note that if there are any fundamental bosons the number of possible states comes out to be strictly infinite.

But then one may come to appreciate this result after all. It means that, given any closed surface, we can represent all that happens inside it by degrees of freedom on this surface itself. This, one may argue, suggests that quantum gravity should be described entirely by a *topological* quantum field theory, in which all physical degrees of freedom can be projected onto the boundary. One Boolean variable per Planckian surface element should suffice. The fact that the total volume inside is irrelevant may be seen as a blessing since it implies that we do not have to worry about the *metric* inside. The inside metric could be so much curved that an entire universe could be squeezed inside our closed surface, regardless how small it is. Now we see that this possibility will not add to the number of allowed states at all. Our result suggests that one should not worry about creating universes inside test tubes.

The same can be said about wormholes. A wormhole with one end sprouting inside our closed volume and its other end somewhere else could connect the inside of our volume to the outside world, thus adding large quantities of possible states. We now believe that this is not allowed in a decent theory of quantum gravity. In a previous publication [8] it was explained why the functional integral describing black holes probably has to be limited to topologically trivial field configurations only. We have detailed ideas concerning the consistency of such a requirement but will not elaborate on this here. So much for wormholes.

We would like to advocate here a somewhat extreme point of view. We suspect that there simply *are* not more degrees of freedom to talk about than the ones one can draw on a surface, as given by eq. (3) The situation can be compared with a hologram of a three dimensional image on a two-dimensional surface. The image is somewhat blurred because of limitations of the hologram technique, but the blurring is small compared to the uncertainties produced by the usual quantum mechanical fluctuations. The details of the hologram on the surface itself are intricate and contain as much information as is allowed by the finiteness of the wavelength of light - read the Planck length.

It is tempting to take the limit where the surface area goes to infinity, and the surface is locally approximately flat. Our variables on the surface then apparently determine all physical events at one side (the black hole side) of the surface. But since the entropy

of a black hole also refers to all physical fields outside the horizon the *same* degrees of freedom determine what happens at this side. Apparently one must conclude that a two-dimensional surface drawn in a three-space can contain all information concerning the entire three-space. In fact, this should hold for *any* two-surface that ranges to infinity. This suggests that physical degrees of freedom in three-space are not independent but, if considered at Planckian scale, they must be infinitely correlated.

If one could determine the equation of motion of these variables on the two-plane one would also possess the remedy for the black hole information paradox. In a Rindler space we could put our information surface at the origin of the Rindler coordinate frame. The transformation rules for our variables under Lorentz transformations would correspond to the Rindler equations of motion, and pure states would evolve into pure states while the spectrum density of the black hole would continue to be the one dictated by its entropy.

The infinite correlations in three-space could also present a new starting point for resolving the Einstein-Rosen-Podolsky paradox. In terms of present day quantum field theoretical degrees of freedom it is not possible to interpret quantum mechanics deterministically. But certain cellular automaton models can give a faint resemblance to quantum behavior. The discreteness of our degrees of freedom on the two-surface strongly remind us of a 2+1 dimensional cellular automaton. But this would constitute a speculation on top of the previous one. There are various technical problems one has to face if such considerations were to be persecuted further.

In cellular automaton models for quantum mechanics the problem is not the Copenhagen interpretation[1]. This comes out quite naturally. The more tantalizing problem is how to understand the stability of our vacuum state[13]. Models can be constructed producing Hamiltonians that can be realistic but for which there are nearly always negative energy eigenstates. The absence of negative energy states in the real world is probably related to the presence of the gravitational force, for which the time coordinate is handled very differently from non-gravitational theories. But natural curvature of space and time is equally difficult to realize in cellular automaton models of this sort[12]; we just suspect that these problems are related but exactly how is not understood.

But there are other problems as well. One of these is the presence of the group of Lorentz transformations and the fact that this symmetry group is non-compact. This is of course at the heart of the black hole horizon problem. There, Lorentz transformations are playing the role of time translations. In any theory where the physical degrees of freedom are discrete it is extremely difficult to reproduce anything resembling Lorentz invariance <sup>1</sup>. As for other invariances such as rotational invariance, one can usually realize

<sup>1</sup> In principle one could think of realizing one of the non-trivial discrete subgroups of the Lorentz group, but in practice this seems to be impossible to reconcile with locality requirements.

symmetry under one of their finite subgroups and then it is not unnatural to suspect that complete invariance will be recovered in the thermodynamic limit as a consequence of renormalization group effects.

The evolution law of the physical degrees of freedom on a 2-surface is another mystery. Ultimately one wishes to recover full invariance with respect to the Poincaré group (which is a precisely defined invariance group for all states that have asymptotic in- and out-states, as the ones used in an  $S$  matrix formalism, but not in quantum cosmology). Now this implies that one not only needs an evolution law for translations in the time direction but also a quite similar looking law for translations in the direction orthogonal to the surface. These two evolution laws should commute with each other in spite of their complete independence. We will show how to reproduce such a feature in cellular automaton models, however, as we will see, requiring these models to possess physically interesting, that is, sufficiently non-trivial interactions, will remain a difficult obstacle.

The question whether models exist in 3+1 dimensions that are such that the data on a two-dimensional surface will determine all observables elsewhere is an extremely intriguing one. For definiteness, let us consider a rectangular lattice in 4-space. Momentarily we will ignore requirements such as Lorentz covariance; it is sufficient to require that signals do not go faster than some limited speed  $c$ . The prototype of our models is a cellular automaton. The data  $f$  on every site on the lattice can be represented by an integer modulo some number  $p$ . This  $p$  will often be taken to be a prime number. The time evolution is defined by some local law. Conventionally, one imposes that the value of  $f(\mathbf{x}, t)$  be a given function of the values of  $f(\mathbf{x}_i, t - 1)$ , where  $\{\mathbf{x}_i\}$  are a finite set of nearest neighbors of  $\mathbf{x}$ .

Now such a model is deterministic in the classical sense. Quantization can be introduced either by allowing  $f(\mathbf{x}_i, t)$  to be operators in Hilbert space, or even more simply, by declaring all states of the cellular automaton to be basis elements of Hilbert space. "Quantization" is then trivial. There are problems with this latter proposal in that the Hamiltonian then does not seem to possess a well-defined ground state. Again, let's not dwell on that.

In order to obtain "dimensional reduction" we will have to postulate a further constraint: the values of  $f$  on a sheet should fix the values elsewhere. This we can realize in principle by also postulating a law of evolution in the  $z$  direction. Since further discussion of this 4 dimensional problem becomes a bit intricate it is illustrative to remove one dimension and treat space as 2 dimensional, space-time as 3 dimensional. Here we will require that the data on a *line* should be sufficient to determine the data elsewhere. Part of the rectangular lattice is depicted in Fig. 1.

Suppose now that on every plaquette of the lattice a relation among the data  $f$  is

imposed. Thus, in the figure there are constraints of the form

$$g_a(f(A), f(B), f(C), f(D)) = 0, \quad (12a)$$

$$g_b(f(E), f(F), f(G), f(H)) = 0, \quad (12b)$$

$$g_c(f(A), f(B), f(F), f(E)) = 0, \quad (12c)$$

$$g_d(f(D), f(C), f(G), f(H)) = 0, \quad (12d)$$

$$g_e(f(A), f(D), f(H), f(E)) = 0, \quad (12e)$$

$$g_f(f(B), f(C), f(G), f(F)) = 0. \quad (12f)$$

We require all these constraints to be such that if in any of these equations three of the four entries are given the fourth will be uniquely determined.

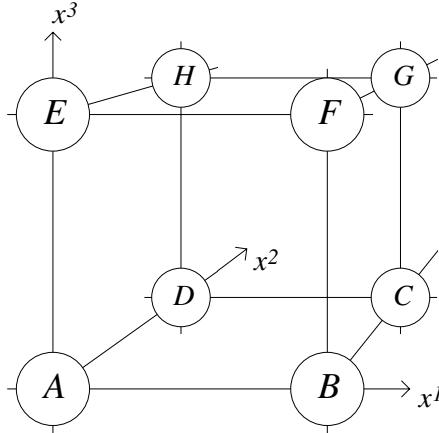


Fig. 1

The lattice sites in  $x^1, x^2, x^3$  coordinates

If we choose the  $x, y, t$  coordinates in one of the principle directions of the lattice the situation becomes a bit singular. It is preferable to define

$$\begin{aligned} t &= x^1 + x^2 + x^3, \\ x &= x^1 - x^2, \\ y &= x^3 - \frac{1}{2}(x^1 + x^2). \end{aligned} \quad (13)$$

if on two nonsecutive  $t$  layers the data are known successive application of eqs (12a-f) will produce the data in all of space-time. But now consider the series of points  $\mathbf{x}(n)$ ,  $n \in \mathbb{Z}$ , defined by

$$\begin{aligned} \mathbf{x}(3n+1) &= \mathbf{x}(3n) + \mathbf{e}^1, \\ \mathbf{x}(3n+2) &= \mathbf{x}(3n+1) - \mathbf{e}^2, \\ \mathbf{x}(3n+3) &= \mathbf{x}(3n+2) + \mathbf{e}^3, \end{aligned} \quad (14)$$

where  $\mathbf{e}^{1,2,3}$  are the three unit vectors of the lattice (the signs in front of the  $\mathbf{e}^i$  can actually be chosen at will). Suppose  $f(\mathbf{x}(n))$  are given for all  $n$ . Successive application of the six identities (12 a–f) then also fixes all data elsewhere.

However, one cannot choose the constraints (12a–f) any way one likes. This is because the data elsewhere will be *over-determined*. Suppose that in the Figure the data are given at  $D$ ,  $H$ ,  $E$  and  $F$ . They form part of a series (14). By applying four of the six equations (12) the other data on the cube are determined. The remaining two must then be satisfied automatically. If these would not be satisfied our system would be constrained further, in such a way that practically no solutions survive at all. The point is that eqs (12a, b) generate translations along the  $x^1x^2$  plane, eqs. (12c, d) translations along the  $x^1x^3$  plane and (12e, f) translations along the  $x^2x^3$  plane. These three translation operators, viewed as operators in Hilbert space, should commute with each other. Only when they are chosen very meticulously these commutation requirements can be met. One can also say that we have translation operators defined on the one dimensional series of data on the points (14). They define translations in the directions  $\mathbf{e}^i \pm \mathbf{e}^j$ , which should all commute with each other. Two linear combinations of these,  $U_1(\delta t)$  and  $U_2(\delta y)$ , should generate *independent* translations in directions orthogonal to the line  $\sigma(\mathbf{e}^1 - \mathbf{e}^2 + \mathbf{e}^3)$  and a third,  $U_3$ , corresponds to a translation in the direction of the line. The three independent translation operators one then has should commute with each other. Commutation with  $U_3$  is usually easy to implement by choosing the evolution not to depend on the coordinate  $n$ , but commutation of  $U_1$  with  $U_2$  is as hard as finding two different yet commuting local Hamiltonians for a quantum system.

Let us phrase the constraints the following way. Consider Fig. 1. One is free to choose  $f(A)$ ,  $f(B)$ ,  $f(D)$  and  $f(E)$ . Then the values of  $f(C)$ ,  $f(F)$  and  $f(H)$  are uniquely determined by eqs (12a, c) and (e). But the value of  $f(G)$  is overdetermined. The three equations (12b, d) and (f) should all yield the same value. Writing the solutions to eqs (12a–f) as

$$f(C) = h_a(f(A), f(B), f(D)) , \quad (15a)$$

$$f(G) = h_b(f(E), f(F), f(H)) , \quad (15b)$$

$$f(F) = h_c(f(A), f(B), f(E)) , \quad (15c)$$

$$f(G) = h_d(f(D), f(C), f(H)) , \quad (15d)$$

$$f(H) = h_e(f(A), f(D), f(E)) , \quad (15e)$$

$$f(G) = h_f(f(B), f(C), f(F)) , \quad (15f)$$

our requirement corresponds to

$$\begin{aligned}
h_b(f(E), h_c(f(A), f(B), f(E)), h_e(f(A), f(D), f(E))) &= \\
h_d(f(D), h_c(f(A), f(B), f(D)), h_e(f(A), f(D), f(E))) &= \quad (16) \\
h_f(f(B), h_c(f(A), f(B), f(D)), h_b(f(A), f(B), f(E))) .
\end{aligned}$$

An easy way to implement these commutation constraints is by choosing the functions  $g_i$  to be *linear in the functions  $f_i$  modulo p* :

$$g_i(\{f_j\}) = \sum_j A_{ij} f_j + B_i \pmod{p}, \quad (17)$$

where the coefficients  $A_{ij}$  must all have an inverse modulo  $p$ . From this one gets three equations for  $f(G)$ :

$$f(G) = K_{1,\alpha} f(A) + K_{2,\alpha} f(B) + K_{3,\alpha} f(D) + K_{4,\alpha} f(E) + K_{5,\alpha} \pmod{p}, \quad \alpha = 1, 2, 3. \quad (18)$$

It is not hard to find sets of coefficients  $A_{ij}$ ,  $B_i$  such that the 10 equations

$$K_{i,1} = K_{i,2} = K_{i,3} \quad (19)$$

are obeyed.

A next step is to attempt to find other realizations of our commutation requirement. Using a computer search program the author generated solutions which at first sight seemed to be quite different, but careful analysis revealed that all solutions found were actually equivalent to the linear ones, eq. (17) after applying permutation operations on the numbers  $f$ . This could be seen as a disappointment because one might argue that a linear relation such as eq. (17) is too trivial to be of much physical interest. It implies that solutions can be superimposed onto each other, as if we were describing only “non-interacting” particles. Our challenge at present is to find any set of plaquette relations that does not allow a superposition procedure to obtain new solutions from old ones (note that superposition here means addition modulo a number, and is quite distinct from quantum mechanical superposition which is always allowed).

Our problem could be seen to simplify a bit by using different lattices. Basically what we want to achieve is a cellular automaton with two commuting but essentially different evolution laws. Suppose we have a triangular lattice in the  $x - y$  direction as well as the  $x - t$  direction, see Fig. 2. We may demand that the data on the  $x$  axis alone should determine all others. This implies that we have a relation  $U_1$  determining how the data

look in the  $y$  direction, and an evolution operator  $U_2$  in the time direction. We have

$$f(F) = U_1(f(A), f(B)), \quad (20a)$$

$$f(G) = U_1(f(B), f(C)), \quad (20b)$$

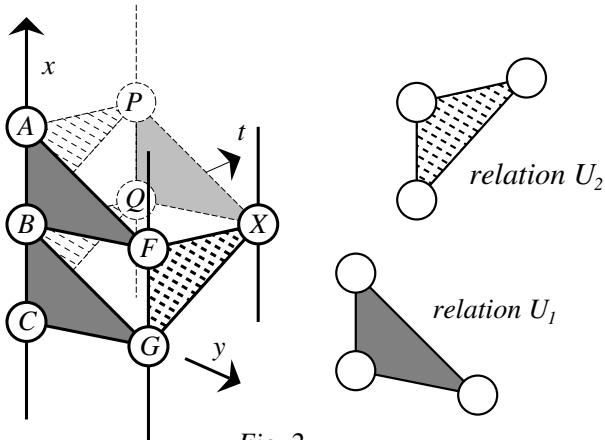
$$f(P) = U_2(f(A), f(B)), \quad (20c)$$

$$f(Q) = U_2(f(B), f(C)), \quad (20d)$$

$$f(X) = U_2(f(F), f(G)), \quad (20e)$$

$$f(X) = U_1(f(P), f(Q)), \quad (20f)$$

and for all choices of  $f(A)$ ,  $f(B)$  and  $f(C)$  the equations (12e) and (12f) should agree. These equations are easier to study than the equations (16). Again there are solutions where  $U_1$  and  $U_2$  are different, providing no direct relations between  $f(F)$  and  $f(P)$  or  $f(G)$  and  $f(Q)$ , but all solutions we found are equivalent to linear ones modulo a prime number  $p$ .



Cellular automaton with two evolution laws on triangular lattices.

In the real world there is one more dimension. By choosing the initial data to be periodic in that one extra dimension the problem reduces to the three dimensional one. Ultimately one would also like to reobtain local Lorentz invariance and coordinate reparametrization invariance. These imply large symmetry groups that are as yet impossible to implement.

The reason why it is interesting to attempt to find an accidental non-trivial solution of the equations is that many cellular automata, defined on different lattice types, can be transformed into each other, as long as the evolution equations act upon nearest neighbors only. So we actually cover a large class of models. If the real world obeys comparable rules it could be equivalent to such an automaton also, though, admittedly, we consider

such a simple structure unlikely. Indeed, one can also devise interaction schemes among discrete variables that are essentially local but *not* equivalent to cellular automata of the kind considered here.

Even though the problems just mentioned are grave and conceptually difficult to disentangle, they do not seem to be insurmountable. Our basic problem is that there seem to be *too many* symmetry requirements and the Hilbert space of physically realizable states seems to be *too small*. The picture we sketched of the 2+1 (or perhaps it is better to say 2+2) dimensional nature of our micro-universe seems to follow from quite general arguments. Rejecting any of these arguments leads to quite different and perhaps equally if not more difficult problems, and one cannot help observing that one's preferences seem to be related to nearly religious prejudices. Besides the author not many other physicists have tried this particular avenue. We advocate its further pursuit.

## References

1. G. 't Hooft, *Nucl. Phys.* **B 342** (1990) 471; *J. Stat. Phys.* **53** (1988) 323.
2. S.W. Hawking, *Phys. Rev.* **D 14** (1976) 2460.
3. S.W. Hawking and G.F.R. Ellis, "The Large Scale Structure of Space-time", Cambridge: Cambridge Univ. Press, 1973.
4. S.W. Hawking, *Phys. Rev.* **D 37** (1988) 904; S. Coleman, *Nucl. Phys.* **B 307** (1988) 864; *ibid.* **B 310** (1988) 643; S.B. Giddings and A. Strominger, *Nucl. Phys.* **B 307** (1988) 854.
5. C.G. Callen, S.B. Giddings, J.A. Harvey and A. Strominger, *Phys. Rev.* **D 45** (1992) R1005.
6. V.P. Frolov and I.G. Novikov, *Phys. Rev.* **D 42** (1990) 1057.
7. T. Dray and G. 't Hooft, *Nucl. Phys.* **B 253** (1985) 173; T. Dray and G. 't Hooft, *Commun. Math. Phys.* **99** (1985) 613; G.'t Hooft, *Nucl. Phys.* **B 335** (1990) 138.
8. C.R. Stephens, G. 't Hooft and B.F. Whiting, "Black hole evaporation without information loss", Utrecht/Gainesville prepr. THU-93/20; UF-RAP-93-11; gr-qc/9310006.
9. G. 't Hooft, "On the Quantization of Space and Time", Proc. of the 4th Seminar on Quantum Gravity, May 25-29, 1987, Moscow, USSR, ed. M.A. Markov et al (World Scientific 1988).
10. G. 't Hooft, *Nucl.Phys.* **B 256** (1985) 727.
11. G. 't Hooft, *Physica Scripta* **T 36** (1991) 247.
12. G. 't Hooft, "A Two-dimensional Model with Discrete General Coordinate-Invariance", in "The Gardener of Eden", Physicalia Magazine, vol **12**, in honour of R. Brout, eds. P. Nicoletopoulos and J. Orloff, Brussels, 1990.
13. G. 't Hooft, K. Isler and S. Kalitzin, *Nucl. Phys.* **B 386** (1992) 495



# STRING THEORY DYNAMICS IN VARIOUS DIMENSIONS

Edward Witten

*School of Natural Sciences, Institute for Advanced Study*

*Olden Lane, Princeton, NJ 08540, USA*

The strong coupling dynamics of string theories in dimension  $d \geq 4$  are studied. It is argued, among other things, that eleven-dimensional supergravity arises as a low energy limit of the ten-dimensional Type IIA superstring, and that a recently conjectured duality between the heterotic string and Type IIA superstrings controls the strong coupling dynamics of the heterotic string in five, six, and seven dimensions and implies  $S$  duality for both heterotic and Type II strings.

March, 1995

## 1. Introduction

Understanding in what terms string theories should really be formulated is one of the basic needs and goals in the subject. Knowing some of the phenomena that can occur for strong coupling – if one can know them without already knowing the good formulation! – may be a clue in this direction. Indeed,  $S$ -duality between weak and strong coupling for the heterotic string in four dimensions (for instance, see [1,2]) really ought to be a clue for a new formulation of string theory.

At present there is very strong evidence for  $S$ -duality in supersymmetric field theories, but the evidence for  $S$ -duality in string theory is much less extensive. One motivation for the present work was to improve this situation.

Another motivation was to try to relate four-dimensional  $S$ -duality to statements or phenomena in more than four dimensions. At first sight, this looks well-nigh implausible since  $S$ -duality between electric and magnetic charge seems to be very special to four dimensions. So we are bound to learn something if we succeed.

Whether or not a version of  $S$ -duality plays a role, one would like to determine the strong coupling behavior of string theories above four dimensions, just as  $S$ -duality – and its conjectured Type II analog, which has been called  $U$ -duality [3] – determines the strong coupling limit after toroidal compactification to four dimensions.<sup>1</sup> One is curious about the phenomena that may arise, and in addition if there is any non-perturbative inconsistency in the higher dimensional string theories (perhaps ultimately leading to an explanation of why we live in four dimensions) it might show up naturally in thinking about the strong coupling behavior.

In fact, in this paper, we will analyze the strong coupling limit of certain string theories in certain dimensions. Many of the phenomena are indeed novel, and many of them are indeed related to dualities. For instance, we will argue in section two that the strong coupling limit of Type IIA supergravity in ten dimensions is eleven-dimensional supergravity! In a sense, this statement gives a rationale for “why” eleven-dimensional

---

<sup>1</sup> By “strong coupling limit” I mean the limit as the string coupling constant goes to infinity keeping fixed (in the sigma model sense) the parameters of the compactification. Compactifications that are not explicitly described or clear from the context will be toroidal.

supergravity exists, much as the interpretation of supergravity theories in  $d \leq 10$  as low energy limits of string theories explains “why” these remarkable theories exist. How eleven-dimensional supergravity fits into the scheme of things has been a puzzle since the theory was first predicted [5] and constructed [6].

Upon toroidal compactification, one can study the strong coupling behavior of the Type II theory in  $d < 10$  using  $U$  duality, as we will do in section three. One can obtain a fairly complete picture, with eleven-dimensional supergravity as the only “surprise.”

Likewise, we will argue in section four that the strong coupling limit of five-dimensional heterotic string theory is Type IIB in six dimensions, while the strong coupling limit of six-dimensional heterotic string theory is Type IIA in six dimensions (in each case with four dimensions as a K3), and the strong coupling limit in seven dimensions involves eleven-dimensional supergravity. These results are based on a relation between the heterotic string and the Type IIA superstring in six dimensions that has been proposed before [3,4]. The novelty in the present paper is to show, for instance, that vexing puzzles about the strong coupling behavior of the heterotic string in five dimensions disappear if one assumes the conjectured relation of the heterotic string to Type IIA in six dimensions. Also we will see – using a mechanism proposed previously in a more abstract setting [7] – that the “string-string duality” between heterotic and Type IIA strings in six dimensions implies  $S$ -duality in four dimensions, so the usual evidence for  $S$ -duality can be cited as evidence for string-string duality.

There remains the question of determining the strong coupling dynamics of the heterotic string above seven dimensions. In this context, there is a curious speculation<sup>2</sup> that the heterotic string in ten dimensions with  $SO(32)$  gauge group might have for its strong coupling limit the  $SO(32)$  Type I theory. In section five, we show that this relation, if valid, straightforwardly determines the strong coupling behavior of the heterotic string in nine and eight dimensions as well as ten, conjecturally completing the description of strong coupling dynamics except for  $E_8 \times E_8$  in ten dimensions.

The possible relations between different theories discussed in this paper should be

---

<sup>2</sup> This idea was considered many years ago by M. B. Green, the present author, and probably others, but not in print as far as I know.

taken together with other, better established relations between different string theories. It follows from  $T$  duality that below ten dimensions the  $E_8 \times E_8$  heterotic string is equivalent to the  $SO(32)$  heterotic string [8,9], and Type IIA is equivalent to Type IIB [10,11]. Combining these statements with the much shakier relations discussed in the present paper, one would have a web of connections between the five string theories and eleven-dimensional supergravity.

After this paper was written and circulated, I learned of a paper [12] that has some overlap with the contents of section two of this paper.

## 2. Type II Superstrings In Ten Dimensions

### 2.1. Type IIB In Ten Dimensions

In this section, we will study the strong coupling dynamics of Type II superstrings in ten dimensions. We start with the easy case, Type IIB. A natural conjecture has already been made by Hull and Townsend [3]. Type IIB supergravity in ten dimensions has an  $SL(2, \mathbf{R})$  symmetry; the conjecture is that an  $SL(2, \mathbf{Z})$  subgroup of this is an exact symmetry of the string theory.<sup>3</sup> This then would relate the strong and weak coupling limits just as  $S$ -duality relates the strong and weak coupling limits of the heterotic string in four dimensions.

This  $SL(2, \mathbf{Z})$  symmetry in ten dimensions, if valid, has powerful implications below ten dimensions. The reason is that in  $d < 10$  dimensions, the Type II theory (Type IIA and Type IIB are equivalent below ten dimensions) is known to have a  $T$ -duality symmetry  $SO(10 - d, 10 - d; \mathbf{Z})$ . This  $T$ -duality group does not commute with the  $SL(2, \mathbf{Z})$  that is already present in ten dimensions, and together they generate the discrete subgroup of the supergravity symmetry group that has been called  $U$ -duality.<sup>4</sup> Thus,  $U$ -duality is true in

<sup>3</sup> For earlier work on the possible role of the non-compact supergravity symmetries in string and membrane theory, see [13].

<sup>4</sup> For instance, in five dimensions,  $T$ -duality is  $SO(5, 5)$  and  $U$ -duality is  $E_6$ . A proper subgroup of  $E_6$  that contains  $SO(5, 5)$  would have to be  $SO(5, 5)$  itself or  $SO(5, 5) \times \mathbf{R}^*$  ( $\mathbf{R}^*$  is the non-compact form of  $U(1)$ ), so when one tries to adjoin to  $SO(5, 5)$  the  $SL(2)$  that was already present in ten dimensions (and contains two generators that map NS-NS states to RR states and so are

every dimension below ten if the  $SL(2, \mathbf{Z})$  of the Type IIB theory holds in ten dimensions.

In the next section we will see that  $U$ -duality controls Type II dynamics below ten dimensions. As  $SL(2, \mathbf{Z})$  also controls Type IIB dynamics in ten dimensions, this fundamental duality between strong and weak coupling controls all Type II dynamics in all dimensions except for the odd case of Type IIA in ten dimensions. But that case will not prove to be a purely isolated exception: the basic phenomenon that we will find in Type IIA in ten dimensions is highly relevant to Type II dynamics below ten dimensions, as we will see in section three. In a way ten-dimensional Type IIA proves to exhibit the essential new phenomenon in the simplest context.

To compare to  $N = 1$  supersymmetric dynamics in four dimensions [14], ten-dimensional Type IIA is somewhat analogous to supersymmetric QCD with  $3N_c/2 > N_f > N_c + 1$ , whose dynamics is controlled by an effective infrared theory that does not make sense at all length scales. The other cases are analogous to the same theory with  $3N_c > N_f > 3N_c/2$ , whose dynamics is controlled by an exact equivalence of theories – conformal fixed points – that make sense at all length scales.

## 2.2. Ramond-Ramond Charges In Ten-Dimensional Type IIA

It is a familiar story to string theorists that the string coupling constant is really the expectation of a field – the dilaton field  $\phi$ . Thus, it can be scaled out of the low energy effective action by shifting the value of the dilaton.

After scaling other fields properly, this idea can be implemented in closed string theories by writing the effective action as  $e^{-2\phi}$  times a function that is invariant under  $\phi \rightarrow \phi + \text{constant}$ . There is, however, an important subtlety here that affects the Type IIA and Type IIB (and Type I) theories. These theories have massless antisymmetric tensor fields that originate in the Ramond-Ramond (RR) sector. If  $A_p$  is such a  $p$ -form field, the natural gauge invariance is  $\delta A_p = d\lambda_{p-1}$ , with  $\lambda_{p-1}$  a  $p - 1$ -form – and no dilaton in the transformation laws. If one scales  $A_p$  by a power of  $e^\phi$ , the gauge transformation law becomes more complicated and less natural.

---

not in  $SO(5, 5)$ ) one automatically generates all of  $E_6$ .

Let us, then, consider the Type IIA theory with the fields normalized in a way that makes the gauge invariance natural. The massless bosonic fields from the (Neveu – Schwarz)<sup>2</sup> or NS-NS sector are the dilaton, the metric tensor  $g_{mn}$ , and the antisymmetric tensor  $B_{mn}$ . From the RR sector, one has a one-form  $A$  and a three form  $A_3$ . We will write the field strengths as  $H = dB$ ,  $F = dA$ , and  $F_4 = dA_3$ ; one also needs  $F'_4 = dA_3 + A \wedge H$ . The bosonic part of the low energy effective action can be written  $I = I_{NS} + I_R$  where  $I_{NS}$  is the part containing NS-NS fields only and  $I_R$  is bilinear in RR fields. One has (in units with  $\alpha' = 1$ )

$$I_{NS} = \frac{1}{2} \int d^{10}x \sqrt{g} e^{-2\phi} \left( R + 4(\nabla\phi)^2 - \frac{1}{12}H^2 \right) \quad (2.1)$$

and

$$I_R = - \int d^{10}x \sqrt{g} \left( \frac{1}{2 \cdot 2!} F^2 + \frac{1}{2 \cdot 4!} F'^2 \right) - \frac{1}{4} \int F_4 \wedge F_4 \wedge B. \quad (2.2)$$

With this way of writing the Lagrangian, the gauge transformation laws of  $A$ ,  $B$ , and  $A_3$  all have the standard, dilaton-independent form  $\delta X = d\Lambda$ , but it is not true that the classical Lagrangian scales with the dilaton like an overall factor of  $e^{-2\phi}$ .

Our interest will focus on the presence of the abelian gauge field  $A$  in the Type IIA theory. The charge  $W$  of this gauge field has the following significance. The Type IIA theory has two supersymmetries in ten dimensions, one of each chirality; call them  $Q_\alpha$  and  $Q'_{\dot{\alpha}}$ . The space-time momentum  $P$  appears in the anticommutators  $\{Q, Q\} \sim \{Q', Q'\} \sim P$ . In the anticommutator of  $Q$  with  $Q'$  it is possible to have a Lorentz-invariant central charge

$$\{Q_\alpha, Q'_{\dot{\alpha}}\} \sim \delta_{\alpha\dot{\alpha}} W. \quad (2.3)$$

To see that such a term does arise, it is enough to consider the interpretation of the Type IIA theory as the low energy limit of eleven-dimensional supergravity, compactified on  $\mathbf{R}^{10} \times \mathbf{S}^1$ . From that point of view, the gauge field  $A$  arises from the components  $g_{m,11}$  of the eleven-dimensional metric tensor,  $W$  is simply the eleventh component of the momentum, and (2.3) is part of the eleven-dimensional supersymmetry algebra.<sup>5</sup>

<sup>5</sup> The relation of the supersymmetry algebra to eleven dimensions leads to the fact that both for the lowest level and even for the first excited level of the Type IIA theory, the states can be

In the usual fashion [17], the central charge (2.3) leads to an inequality between the mass  $M$  of a particle and the value of  $W$ :

$$M \geq c_0 |W|, \quad (2.4)$$

with  $c_0$  a “constant,” that is a function only of the string coupling constant  $\lambda = e^\phi$ , and independent of which particle is considered. The precise constant with which  $W$  appears in (2.3) or (2.4) can be worked out using the low energy supergravity (there is no need to worry about stringy corrections as the discussion is controlled by the leading terms in the low energy effective action, and these are uniquely determined by supersymmetry). We will work this out at the end of this section by a simple scaling argument starting with eleven-dimensional supergravity. For now, suffice it to say that the  $\lambda$  dependence of the inequality is actually

$$M \geq \frac{c_1}{\lambda} |W| \quad (2.5)$$

with  $c_1$  an absolute constant. States for which the inequality is saturated – we will call them BPS-saturated states by analogy with certain magnetic monopoles in four dimensions – are in “small” supermultiplets with  $2^8$  states, while generic supermultiplets have  $2^{16}$  states.

In the elementary string spectrum,  $W$  is identically zero. Indeed, as  $A$  originates in the RR sector,  $W$  would have had to be a rather exotic charge mapping NS-NS to RR states. However, there is no problem in finding classical black hole solutions carrying the  $W$  charge (or any other gauge charge, in any dimension). It was proposed by Hull and Townsend [3] that quantum particles carrying RR charges arise by quantization of such black holes. Recall that, in any dimension, charged black holes obey an inequality  $GM^2 \geq \text{const} \cdot W^2$  ( $G$ ,  $M$ , and  $W$  are Newton’s constant and the black hole mass and charge); with  $G \sim \lambda^2$ , this inequality has the same structure as (2.5). These two inequalities actually correspond in the sense that an extreme black hole, with the minimum mass for given charge, is invariant under some supersymmetry [18] and so should correspond upon quantization to a “small” supermultiplet saturating the inequality (2.5).

---

arranged in eleven-dimensional Lorentz multiplets [15]. If this would persist at higher levels, it might be related to the idea that will be developed below. It would also be interesting to look for possible eleven-dimensional traces in the superspace formulation [16].

To proceed, then, I will assume that there are in the theory BPS-saturated particles with  $W \neq 0$ . This assumption can be justified as follows. Hull and Townsend actually showed that upon toroidally compactifying to less than ten dimensions, the assumption follows from  $U$ -duality. In toroidal compactification, the radii of the circles upon which one compactifies can be arbitrarily big. That being so, it is implausible to have BPS-saturated states of  $W \neq 0$  below ten dimensions unless they exist in ten dimensions; that is, if the smallest mass of a  $W$ -bearing state in ten dimensions were strictly bigger than  $c|W|/\lambda$ , then this would remain true after compactification on a sufficiently big torus.

If the ten-dimensional theory has BPS-saturated states of  $W \neq 0$ , then what values of  $W$  occur? A continuum of values of  $W$  would seem pathological. A discrete spectrum is more reasonable. If so, the quantum of  $W$  must be independent of the string coupling “constant”  $\lambda$ . The reason is that  $\lambda$  is not really a “constant” but the expectation value of the dilaton field  $\phi$ . If the quantum of  $W$  were to depend on the value of  $\phi$ , then the value of the electric charge  $W$  of a particle would change in a process in which  $\phi$  changes (that is, a process in which  $\phi$  changes in a large region of space containing the given particle); this would violate conservation of  $W$ .

The argument just stated involves a hidden assumption that will now be made explicit. The canonical action for a Maxwell field is

$$\frac{1}{4e^2} \int d^n x \sqrt{g} F^2. \quad (2.6)$$

Comparing to (2.2), we see that in the case under discussion the effective value of  $e$  is independent of  $\phi$ , and this is why the charge of a hypothetical charged particle is independent of  $\phi$ . If the action were

$$\frac{1}{4} \int d^n x \sqrt{g} e^{\gamma\phi} F^2 \quad (2.7)$$

for some non-zero  $\gamma$ , then the current density would equal (from the equations of motion of  $A$ )  $J_m = \partial^n (e^{\gamma\phi} F_{mn})$ . In a process in which  $\phi$  changes in a large region of space containing a charge, there could be a current inflow proportional to  $\nabla\phi \cdot F$ , and the charge would in fact change. Thus, it is really the  $\phi$ -independence of the kinetic energy of the RR

fields that leads to the statement that the values of  $W$  must be independent of the string coupling constant and that the masses of charged fields scale as  $\lambda^{-1}$ .

Since the classical extreme black hole solution has arbitrary charge  $W$  (which can be scaled out of the solution in an elementary fashion), one would expect, if BPS-saturated charged particles do arise from quantization of extreme black holes, that they should possess every allowed charge. Thus, we expect BPS-saturated extreme black holes of mass

$$M = \frac{c|n|}{\lambda}, \quad (2.8)$$

where  $n$  is an arbitrary integer, and, because of the unknown value of the quantum of electric charge,  $c$  may differ from  $c_1$  in (2.5).

Apart from anything else that follows, the existence of particles with masses of order  $1/\lambda$ , as opposed to the more usual  $1/\lambda^2$  for solitons, is important in itself. It almost certainly means that the string perturbation expansion – which is an expansion in powers of  $\lambda^2$  – will have non-perturbative corrections of order  $\exp(-1/\lambda)$ , in contrast to the more usual  $\exp(-1/\lambda^2)$ <sup>6</sup>. The occurrence of such terms has been guessed by analogy with matrix models [20].

The fact that the masses of RR charges diverge as  $\lambda \rightarrow 0$  – though only as  $1/\lambda$  – is important for self-consistency. It means that these states disappear from the spectrum as  $\lambda \rightarrow 0$ , which is why one does not see them as elementary string states.

### 2.3. Consequences For Dynamics

Now we will explore the consequences for dynamics of the existence of these charged particles.

The mass formula (2.8) shows that, when the string theory is weakly coupled, the RR charges are very heavy. But if we are bold enough to follow the formula into strong coupling, then for  $\lambda \rightarrow \infty$ , these particles go to zero mass. This may seem daring, but the familiar argument based on the “smallness” of the multiplets would appear to show that the formula (2.8) is exact and therefore can be used even for strong coupling. In

---

<sup>6</sup> If there are particles of mass  $1/\lambda$ , then loops of those particles should give effects of order  $e^{-1/\lambda}$ , while loops of conventional solitons, with masses  $1/\lambda^2$ , would be of order  $\exp(-1/\lambda^2)$ .

four dimensions, extrapolation of analogous mass formulas to strong coupling has been extremely successful, starting with the original work of Montonen and Olive that led to the idea of  $S$ -duality. (In four-dimensional  $N = 2$  theories, such mass formulas generally fail to be exact [21] because of quantum corrections to the low energy effective action. For  $N = 4$  in four dimensions, or for Type IIA supergravity in ten dimensions, the relevant, leading terms in the low energy action are uniquely determined by supersymmetry.)

So for strong coupling, we imagine a world in which there are supermultiplets of mass  $M = c|n|/\lambda$  for every  $\lambda$ . These multiplets necessarily contain particles of spin at least two, as every supermultiplet in Type IIA supergravity in ten dimensions has such states. (Multiplets that do not saturate the mass inequality contain states of spin  $\geq 4$ .) Rotation-invariance of the classical extreme black hole solution suggests<sup>7</sup> (as does  $U$ -duality) that the BPS-saturated multiplets are indeed in this multiplet of minimum spin.

Thus, for  $\lambda \rightarrow \infty$  we have light, charged fields of spin two. (That is, they are charged with respect to the ten-dimensional gauge field  $A$ .) Moreover, there are infinitely many of these. This certainly does not correspond to a local field theory in ten dimensions. What kind of theory will reproduce this spectrum of low-lying states? One is tempted to think of a string theory or Kaluza-Klein theory that has an infinite tower of excitations. The only other option, really, is to assume that the strong coupling limit is a sort of theory that we do not know about at all at present.

One can contemplate the possibility that the strong coupling limit is some sort of a string theory with the dual string scale being of order  $1/\lambda$ , so that the charged multiplets under discussion are some of the elementary string states. There are two reasons that this approach does not seem promising: (i) there is no known string theory with the right properties (one needs Type IIA supersymmetry in ten dimensions, with charged string states coupling to the abelian gauge field in the gravitational multiplet); (ii) we do not have evidence for a stringy exponential proliferation of light states as  $\lambda \rightarrow \infty$ , but only for

<sup>7</sup> Were the classical solution not rotationally invariant, then upon quantizing it one would obtain a band of states of varying angular momentum. One would then not expect to saturate the mass inequality of an extreme black hole without taking into account the angular momentum.

a single supermultiplet for each integer  $n$ , with mass  $\sim |n|$ .

Though meager compared to a string spectrum, the spectrum we want to reproduce is just about right for a Kaluza-Klein theory. Suppose that in the region of large  $\lambda$ , one should think of the theory not as a theory on  $\mathbf{R}^{10}$  but as a theory on  $\mathbf{R}^{10} \times \mathbf{S}^1$ . Such a theory will have a “charge” coming from the rotations of  $\mathbf{S}^1$ . Suppose that the radius  $r(\lambda)$  of the  $\mathbf{S}^1$  scales as  $1/\lambda$  (provided that distances are measured using the “string” metric that appears in (2.1) – one could always make a Weyl rescaling). Then for large  $\lambda$ , each massless field in the eleven-dimensional theory will give, in ten dimensions, for each integer  $n$  a single field of charge  $n$  and mass  $\sim |n|\lambda$ . This is precisely the sort of spectrum that we want.

So we need an eleven-dimensional field theory whose fields are in one-to-one correspondence with the fields of the Type IIA theory in ten dimensions. Happily, there is one: eleven-dimensional supergravity! So we are led to the strange idea that eleven-dimensional supergravity may govern the strong coupling behavior of the Type IIA superstring in ten dimensions.

Let us discuss a little more precisely how this would work. The dimensional reduction of eleven-dimensional supergravity to ten dimensions including the massive states has been discussed in some detail (for example, see [22]). Here we will be very schematic, just to touch on the points that are most essential. The bosonic fields in eleven-dimensional supergravity are the metric  $G_{MN}$  and a three-form  $A_3$ . The bosonic part of the action is

$$I = \frac{1}{2} \int d^{11}x \sqrt{G} (R + |dA_3|^2) + \int A_3 \wedge dA_3 \wedge dA_3. \quad (2.9)$$

Now we reduce to 10 dimensions, taking the eleventh dimensions to be a circle of radius  $e^\gamma$ . That is, we take the eleven-dimensional metric to be  $ds^2 = G_{mn}^{10} dx^m dx^n + e^{2\gamma} (dx^{11} - A_m dx^m)^2$  to describe a ten-dimensional metric  $G^{10}$  along with a vector  $A$  and scalar  $\gamma$ ; meanwhile  $A_3$  reduces to a three-form which we still call  $A_3$ , and a two-form  $B$  (the part of the original  $A_3$  with one index equal to 11). Just for the massless fields, the bosonic part of the action becomes roughly

$$I = \frac{1}{2} \int d^{10}x \sqrt{G^{10}} (e^\gamma (R + |\nabla\gamma|^2 + |dA_3|^2) + e^{3\gamma} |dA|^2 + e^{-\gamma} |dB|^2) + \dots \quad (2.10)$$

This formula, like others below, is very rough and is only intended to exhibit the powers of  $e^\gamma$ . The point in its derivation is that, for example, the part of  $A_3$  that does not have an index equal to “11” has a kinetic energy proportional to  $e^\gamma$ , while the part with such an index has a kinetic energy proportional to  $e^{-\gamma}$ .

The powers of  $e^\gamma$  in (2.10) do not, at first sight, appear to agree with those in (2.1). To bring them in agreement, we make a Weyl rescaling by writing  $G^{10} = e^{-\gamma}g$ . Then in terms of the new ten-dimensional metric  $g$ , we have

$$I = \frac{1}{2} \int d^{10}x \sqrt{g} (e^{-3\gamma} (R + |\nabla\gamma|^2 + |dB|^2) + |dA|^2 + |dA_3|^2 + \dots). \quad (2.11)$$

We see that (2.11) now does agree with (2.1) if

$$e^{-2\phi} = e^{-3\gamma}. \quad (2.12)$$

In the original eleven-dimensional metric, the radius of the circle is  $r(\lambda) = e^\gamma$ , but now, relating  $\gamma$  to the dilaton string coupling constant via (2.12), we can write

$$r(\lambda) = e^{\frac{2\phi}{3}} = \lambda^{2/3}. \quad (2.13)$$

The masses of Kaluza-Klein modes of the eleven-dimensional theory are of order  $1/r(\lambda)$  when measured in the metric  $G^{10}$ , but in the metric  $g$  they are of order

$$\frac{e^{-\gamma/2}}{r(\lambda)} \sim \lambda^{-1}. \quad (2.14)$$

Manipulations similar to what we have just seen will be made many times in this paper.

Here are the salient points:

- (1) The radius of the circle *grows* by the formula (2.13) as  $\lambda \rightarrow \infty$ . This is important for self-consistency; it means that when  $\lambda$  is large the eleven-dimensional theory is weakly coupled at its compactification scale. Otherwise the discussion in terms of eleven-dimensional field theory would not make sense, and we would not know how to improve on it. As it is, our proposal reduces the strongly coupled Type IIA superstring to a field theory that is weakly coupled at the scale of the low-lying excitations, so we get an effective determination of the strong coupling behavior.

(2) The mass of a particle of charge  $n$ , measured in the string metric  $g$  in the effective ten-dimensional world, is of order  $|n|/\lambda$  from (2.14). This is the dependence on  $\lambda$  claimed in (2.5), which we have now in essence derived: the dependence of the central charge on  $\phi$  is uniquely determined by the low energy supersymmetry, so by deriving this dependence in a Type IIA supergravity theory that comes by Kaluza-Klein reduction from eleven dimensions, we have derived it in general.

So far, the case for relating the strong coupling limit of Type IIA superstrings to eleven-dimensional supergravity consists of the fact that this enables us to make sense of the otherwise puzzling dynamics of the BPS-saturated states and that point (1) above worked out correctly, which was not obvious *a priori*. The case will hopefully get much stronger in the next section when we extend the analysis to work below ten dimensions and incorporate  $U$ -duality, and in section four when we look at the heterotic string in seven dimensions. In fact, the most startling aspect of relating strong coupling string dynamics to eleven-dimensional supergravity is the Lorentz invariance that this implies between the eleventh dimension and the original ten. Both in section three and in section four, we will see remnants of this underlying Lorentz invariance.

### 3. Type II Dynamics Below Ten Dimensions

#### 3.1. $U$ -Duality And Dynamics

In this section, we consider Type II superstrings toroidally compactified to  $d < 10$  dimensions, with the aim of understanding the strong coupling dynamics, that is the behavior when some parameters, possibly including the string coupling constant, are taken to extreme values.

The strong coupling behaviors of Type IIA and Type IIB seem to be completely different in ten dimensions, as we have seen. Upon toroidal compactification below ten dimensions, the two theories are equivalent under  $T$ -duality [10,11], and so can be considered together. We will call the low energy supergravity theory arising from this compactification Type II supergravity in  $d$  dimensions.

The basic tool in the analysis is  $U$ -duality. Type II supergravity in  $d$  dimensions

has a moduli space of vacua of the form  $G/K$ , where  $G$  is a non-compact connected Lie group (which depends on  $d$ ) and  $K$  is a compact subgroup, generally a maximal compact subgroup of  $G$ .  $G$  is an exact symmetry of the supergravity theory. There are also  $U(1)$  gauge bosons, whose charges transform as a representation of  $G$ .<sup>8</sup> The structure was originally found by dimensional reduction from eleven dimensions [23].

In the string theory realization, the moduli space of vacua remains  $G/K$  since this is forced by the low energy supergravity. Some of the Goldstone bosons parametrizing  $G/K$  come from the NS-NS sector and some from the RR sector. The same is true of the gauge bosons. In string theory, the gauge bosons that come from the NS-NS sector couple to charged states in the elementary string spectrum. It is therefore impossible for  $G$  to be an exact symmetry of the string theory – it would not preserve the lattice of charges. The  $U$ -duality conjecture says that an integral form of  $G$ , call it  $G(\mathbf{Z})$ , is a symmetry of the string theory. If so, then as the NS-NS gauge bosons couple to BPS-saturated charges, the same must be true of the RR gauge bosons – though the charges in question do not appear in the elementary string spectrum. The existence of such RR charges was our main assumption in the last section; we see that this assumption is essentially a consequence of  $U$ -duality.

The BPS saturated states are governed by an exact mass formula – which will be described later in some detail – which shows how some of them become massless when one approaches various limits in the moduli space of vacua. Our main dynamical assumption is that the smallest mass scale appearing in the mass formula is always the smallest mass scale in the theory.

We assume that at a generic point in  $G/K$ , the only massless states are those in the supergravity multiplet. There is then nothing to say about the dynamics: the infrared behavior is that of  $d$  dimensional Type II supergravity. There remains the question of

<sup>8</sup> To make a  $G$ -invariant theory on  $G/K$ , the matter fields in general must be in representations of the unbroken symmetry group  $K$ . Matter fields that are in representations of  $K$  that do not extend to representations of  $G$  are sections of some homogeneous vector bundles over  $G/K$  with non-zero curvature. The potential existence of an integer lattice of charges forces the gauge bosons to be sections instead of a flat bundle, and that is why they are in a representation of  $G$  and not only of  $K$ .

what happens when one takes various limits in  $G/K$  – for instance, limits that correspond to weak coupling or large radius or (more mysteriously) strong coupling or very strong excitation of RR scalars. We will take the signal that something interesting is happening to be that the mass formula predicts that some states are going to zero mass. When this occurs, we will try to determine what is the dynamics of the light states, in whatever limit is under discussion.

We will get a complete answer, in the sense that for every degeneration of the Type II superstring in  $d$  dimensions, there is a natural candidate for the dynamics. In fact, there are basically only two kinds of degeneration; one involves weakly coupled string theory, and the other involves weakly coupled eleven-dimensional supergravity. In one kind of degeneration, one sees toroidal compactification of a Type II superstring from ten to  $d$  dimensions; the degeneration consists of the fact that the string coupling constant is going to zero. (The parameters of the torus are remaining fixed.) In the other degeneration one sees toroidal compactification of eleven-dimensional supergravity from eleven to  $d$  dimensions; the degeneration consists of the fact that the radius of the torus is going to infinity so that again the coupling constant at the compactification scale is going to zero.<sup>9</sup> (These are actually the degenerations that produce maximal sets of massless particles; others correspond to perturbations of these.)

Thus, with our hypotheses, one gets a complete control on the dynamics, including strong coupling. Every limit which one might have been tempted to describe as “strong coupling” actually has a weakly coupled description in the appropriate variables. The ability to get this consistent picture can be taken as evidence that the hypotheses are true, that  $U$ -duality is valid, and that eleven-dimensional supergravity plays the role in the dynamics that was claimed in section two.

It may seem unexpected that weakly coupled string theory appears in this analysis as a “degeneration,” where some particles go to zero mass, so let me explain this situation. For  $d < 9$ ,  $G$  is semi-simple, and the dilaton is unified with other scalars. The “string” version

<sup>9</sup> It is only in the eleven-dimensional description that the radius is going to infinity. In the ten-dimensional string theory description, the radius is fixed but the string coupling constant is going to infinity.

of the low energy effective action, in which the dilaton is singled out in the gravitational kinetic energy

$$\int d^d x \sqrt{g} e^{-2\phi} R \quad (3.1)$$

is unnatural for exhibiting such a symmetry. The  $G$ -invariant metric is the one obtained by a Weyl transformation that removes the  $e^{-2\phi}$  from the gravitational kinetic energy. The transformation in question is of course the change of variables  $g = e^{4\phi/(d-2)} g'$ , with  $g'$  the new metric. This transformation multiplies masses by  $e^{2\phi/(d-2)}$ , that is, by

$$w_d = \lambda^{2/(d-2)} \quad (3.2)$$

(with  $\lambda$  the string coupling constant). Thus, while elementary string states have masses of order one with respect to the string metric, their masses are of order  $\lambda^{2/(d-2)}$  in the natural units for discussions of  $U$ -duality. So, from this point of view, the region of weakly coupled string theory is a “degeneration” in which some masses go to zero.

It is amusing to consider that, in a world in which supergravity was known and string theory unknown, the following discussion might have been carried out, with a view to determining the strong coupling limit of a hypothetical consistent theory related to Type II supergravity. The string theory degeneration might then have been found, giving a clue to the existence of this theory. Similarly, the strong coupling analysis that we are about to perform might *a priori* have uncovered new theories beyond string theory and eleven-dimensional supergravity, but this will not be the case.

### 3.2. The Nature Of Infinity

It is useful to first explain – without specific computations – why NS-NS (rather than RR) moduli play the primary role.

We are interested in understanding what particles become light – and how they interact – when one goes to infinity in the moduli space  $G(\mathbf{Z}) \backslash G/K$ . The discussion is simplified by the fact that the groups  $G$  that arise in supergravity are the maximally split forms of the corresponding Lie groups. This simply means that they contain a maximal abelian subgroup  $A$  which is a product of copies of  $\mathbf{R}^*$  (rather than  $U(1)$ ). <sup>10</sup>

---

<sup>10</sup> Algebraists call  $A$  a “maximal torus,” and  $T$  would be the standard name, but I will avoid

For instance, in six dimensions  $G = SO(5, 5)$ , with rank 5. One can think of  $G$  as the orthogonal group acting on the sum of five copies of a two dimensional real vector space  $H$  endowed with quadratic form

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \quad (3.3)$$

Then a maximal abelian subgroup of  $G$  is the space of matrices looking like a sum of five  $2 \times 2$  blocks, of the form

$$\begin{pmatrix} e^{\lambda_i} & 0 \\ 0 & e^{-\lambda_i} \end{pmatrix} \quad (3.4)$$

for some  $\lambda_i$ . This group is of the form  $(\mathbf{R}^*)^5$ . Likewise, the integral forms arising in  $T$  and  $U$  duality are the maximally split forms over  $\mathbf{Z}$ ; for instance the  $T$ -duality group upon compactification to  $10 - d$  dimensions is the group of integral matrices preserving a quadratic form which is the sum of  $d$  copies of (3.3). This group is sometimes called  $SO(d, d; \mathbf{Z})$ .

With the understanding that  $G$  and  $G(\mathbf{Z})$  are the maximally split forms, the structure of infinity in  $G(\mathbf{Z}) \backslash G/K$  is particularly simple. A fundamental domain in  $G(\mathbf{Z}) \backslash G/K$  consists of group elements of the form  $g = tu$ , where the notation is as follows.  $u$  runs over a *compact* subset  $U$  of the space of generalized upper triangular matrices; compactness of  $U$  means that motion in  $U$  is irrelevant in classifying the possible ways to “go to infinity.”  $t$  runs over  $A/W$  where  $A$  was described above, and  $W$  is the Weyl group.

Thus, one can really only go to infinity in the  $A$  direction, and moreover, because of dividing by  $W$ , one only has to consider going to infinity in a “positive” direction.

Actually,  $A$  has a very simple physical interpretation. Consider the special case of compactification from 10 to  $10 - d$  dimensions on an orthogonal product of circles  $\mathbf{S}_i^1$  of radius  $r_i$ . Then  $G$  has rank  $d + 1$ , so  $A$  is a product of  $d + 1$   $\mathbf{R}^*$ ’s.  $d$  copies of  $\mathbf{R}^*$  act by rescaling the  $r_i$  (making up a maximal abelian subgroup of the  $T$ -duality group  $SO(d, d)$ ), and the last one rescales the string coupling constant. So in particular, with this choice of  $A$ , if one starts at a point in moduli space at which the RR fields are all zero, they remain zero under the action of  $A$ .

---

this terminology because (i) calling  $(\mathbf{R}^*)^n$  a “torus” might be confusing, especially in the present context in which there are so many other tori; (ii) in the present problem the letter  $T$  is reserved for the  $T$ -duality group.

Thus, one can probe all possible directions at infinity without exciting the RR fields; directions in which some RR fields go to infinity are equivalent to directions in which one only goes to infinity via NS-NS fields. Moreover, by the description of  $A$  just given, going to infinity in NS-NS directions can be understood to mean just taking the string coupling constant and the radial parameters of the compactification to zero or infinity.

### 3.3. The Central Charges And Their Role

Let us now review precisely why it is possible to predict particle masses from  $U$ -duality. The unbroken subgroup  $K$  of the supergravity symmetry group  $G$  is realized in Type II supergravity as an  $R$ -symmetry group; that is, it acts non-trivially on the supersymmetries.  $K$  therefore acts on the central charges in the supersymmetry algebra. The scalar fields parametrizing the coset space  $G/K$  enable one to write a  $G$ -invariant formula for the central charges (which are a representation of  $K$ ) of the gauge bosons (which are a representation of  $G$ ). For most values of  $d$ , the formula is uniquely determined, up to a multiplicative constant, by  $G$ -invariance, so the analysis does not require many details of supergravity. That is fortunate as not all the details we need have been worked out in the literature, though many can be found in [24].

For example, let us recall (following [3]) the situation in  $d = 4$ . The  $T$ -duality group is  $SO(6, 6)$ , and  $S$ -duality would be  $SL(2)$  (acting on the axion-dilaton system and exchanging electric and magnetic charge).  $SO(6, 6) \times SL(2)$  is a maximal subgroup of the  $U$ -duality group which is  $G = E_7$  (in its non-compact, maximally split form) and has  $K = SU(8)$  as a maximal compact subgroup.

Toroidal compactification from ten to four dimensions produces in the NS-NS sector twelve gauge bosons coupling to string momentum and winding states, and transforming in the twelve-dimensional representation of  $SO(6, 6)$ . The electric and magnetic charges coupling to any one of these gauge bosons transform as a doublet of  $SL(2)$ , so altogether the NS-NS sector generates a total of 24 gauge charges, transforming as  $(\mathbf{12}, \mathbf{2})$  of  $SO(6, 6) \times SL(2)$ .

From the RR sector, meanwhile, one gets 16 vectors. (For instance, in Type IIA, the vector of the ten-dimensional RR sector gives 1 vector in four dimensions, and the

three-form gives  $6 \cdot 5/2 = 15$ .) These 16 states give a total of  $16 \cdot 2 = 32$  electric and magnetic charges, which can be argued to transform in an irreducible spinor representation of  $SO(6, 6)$  (of positive or negative chirality for Type IIA or Type IIB), while being  $SL(2)$  singlet. The fact that these states are  $SL(2)$  singlets means that there is no natural way to say which of the RR charges are electric and which are magnetic. Altogether, there are  $24 + 32 = 56$  gauge charges, transforming as

$$(\mathbf{12}, \mathbf{2}) \oplus (\mathbf{32}, \mathbf{1}) \quad (3.5)$$

under  $SO(6, 6) \times SL(2)$ ; this is the decomposition of the irreducible **56** of  $E_7$ . Let us call the space of these charges  $V$ .

The four-dimensional theory has  $N = 8$  supersymmetry; thus there are eight positive-chirality supercharges  $Q_\alpha^i$ ,  $i = 1 \dots 8$ , transforming in the **8** of  $K = SU(8)$ . The central charges, arising in the formula

$$\{Q_\alpha^i, Q_\beta^j\} = \epsilon_{\alpha\beta} Z^{ij}, \quad (3.6)$$

therefore transform as the second rank antisymmetric tensor of  $SU(8)$ , the **28**: this representation has complex dimension 28 or real dimension 56. Denote the space of  $Z^{ij}$ 's as  $W$ .

Indeed, the **56** of  $E_7$ , when restricted to  $SU(8)$ , coincides with the **28**, regarded as a 56-dimensional real representation. (Equivalently, the **56** of  $E_7$  when complexified decomposes as **28**  $\oplus$  **28** of  $SU(8)$ .) There is of course a natural,  $SU(8)$ -invariant metric on  $W$ . As the **56** is a pseudoreal rather than real representation of  $E_7$ , there is no  $E_7$ -invariant metric on  $V$ . However, as  $V$  and  $W$  coincide when regarded as representations of  $SU(8)$ , one can pick an embedding of  $SU(8)$  in  $E_7$  and then define an  $SU(8)$ -covariant map  $T : V \rightarrow W$  which determines a metric on  $V$ .

There is no reason to pick one embedding rather than another, and indeed the space of vacua  $E_7/SU(8)$  of the low energy supergravity theory can be interpreted as the space of all  $SU(8)$  subgroups of  $E_7$ . Given  $g \in E_7$ , we can replace  $T : V \rightarrow W$  by

$$T_g = Tg^{-1}. \quad (3.7)$$

This is not invariant under  $g \rightarrow gk$ , with  $k \in SU(8)$ , but it is so invariant up to an  $SU(8)$  transformation of  $W$ . So let  $\psi \in V$  be a vector of gauge charges of some string state. Then

$$\psi \rightarrow Z(\psi) = T_g \psi \quad (3.8)$$

gives a vector in  $W$ , representing the central charges of  $\psi$ . The map from “states”  $\psi$  to central charges  $Z(\psi)$  is manifestly  $E_7$ -invariant, that is invariant under

$$\begin{aligned} \psi &\rightarrow g'\psi \\ g &\rightarrow g'g. \end{aligned} \quad (3.9)$$

Also, under  $g \rightarrow gk$ , with  $k \in SU(8)$ ,  $Z$  transforms to  $Tk^{-1}T^{-1}Z$ , that is, it transforms by a “local  $SU(8)$  transformation” that does not affect the norm of the central charge. The formula (3.8) is, up to a constant multiple, the only formula with these properties, so it is the one that must come from the supergravity or superstring theory.

In supersymmetric theories with central charges, there is an inequality between the mass of a state and the central charge. For elementary string winding states and their partners under  $U$ -duality, the inequality is  $M \geq |Z|$ . (More generally, the inequality is roughly that  $M$  is equal to or greater than the largest eigenvalue of  $Z$ ; for a description of stringy black holes with more than one eigenvalue, see [19]. Elementary string states have only one eigenvalue.)

So far, we have not mentioned the integrality of the gauge charges. Actually, states carrying the 56 gauge charges only populate a lattice  $V_{\mathbf{Z}} \subset V$ . If  $U$ -duality is true, then each lattice point related by  $U$ -duality to the gauge charges of an elementary string state represents the charges of a supermultiplet of mass  $|Z(\psi)|$ .

As an example of the use of this formalism, let us keep a promise made in section two and give an alternative deduction, assuming  $U$ -duality, of the important statement that the masses of states carrying RR charges are (in string units) of order  $1/\lambda$ .<sup>11</sup> Starting from any given vacuum, consider the one-parameter family of vacua determined by the following one-parameter subgroup of  $SO(6, 6) \times SL(2)$ : we take the identity in  $SO(6, 6)$

<sup>11</sup> The following argument was pointed out in parallel by C. Hull.

(so that the parameters of the toroidal compactification are constant) times

$$g_t = \begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix} \quad (3.10)$$

in  $SL(2)$  (so as to vary the string coupling constant). We work here in a basis in which the “top” component is electric and the “bottom” component is magnetic.

Using the mass formula  $M(\psi) = |Z(\psi)| = |Tg^{-1}\psi|$ , the  $t$  dependence of the mass of a state comes entirely from the  $g$  action on the state. The NS-NS states, as they are in a doublet of  $SL(2)$ , have “electric” components whose masses scale as  $e^{-t}$  and “magnetic” components with masses of  $e^t$ . On the other hand, as the RR states are  $SL(2)$  singlets, the mass formula immediately implies that their masses are independent of  $t$ .

These are really the masses in the  $U$ -dual “Einstein” metric. Making a Weyl transformation to the “string” basis in which the electric NS-NS states (which are elementary string states) have masses of order one, the masses are as follows: electric NS-NS,  $M \sim 1$ ; magnetic NS-NS,  $M \sim e^{2t}$ ; RR,  $M \sim e^t$ . But since we know that the magnetic NS-NS states (being fairly conventional solitons) have masses of order  $1/\lambda^2$ , we identify  $e^t = 1/\lambda$  (a formula one could also get from the low energy supergravity); hence the RR masses are of order  $1/\lambda$  as claimed.<sup>12</sup>

The basic properties described above hold in any dimension above three. (In nine dimensions, some extra care is needed because the  $U$ -duality group is not semi-simple.) In three dimensions, new phenomena, which we will not try to unravel, appear because vectors are dual to scalars and charges are confined (for some of the relevant material, see [25]).

### 3.4. Analysis Of Dynamics

We now want to justify the claims made at the beginning of this section about the strong coupling dynamics.

---

<sup>12</sup> We made this deduction here in four dimensions, but it could be made, using  $U$ -duality, in other dimensions as well. Outside of four dimensions, instead of using the known mass scale of magnetic monopoles to fix the relation between  $t$  and  $\lambda$ , one could use the known Weyl transformation (3.2) between the string and  $U$ -dual mass scales.

To do this, we will analyze limits of the theory in which some of the BPS-saturated particles go to zero mass. Actually, for each way of going to infinity, we will look only at the particles whose masses goes to zero as fast as possible. We will loosely call these the particles that are massless at infinity.

Also, we really want to find the “maximal” degenerations, which produce maximal sets of such massless particles; a set of massless particles, produced by going to infinity in some direction, is maximal if there would be no way of going to infinity such that those particles would become massless together with others. A degeneration (i.e., a path to infinity) that produces a non-maximal set of massless particles should be understood as a perturbation of a maximal degeneration. (In field theory, such perturbations, which partly lift the degeneracy of the massless particles, are called perturbations by relevant operators.) We will actually also check a few non-maximal degenerations, just to make sure that we understand their physical interpretation.

To justify our claims, we should show that in any dimension  $d$ , there are only two maximal degenerations, which correspond to toroidal compactification of weakly coupled ten-dimensional string theory and to toroidal compactification of eleven-dimensional supergravity, respectively. The analysis is in fact very similar in spirit for any  $d$ , but the details of the group theory are easier for some values of  $d$  than others. I will first explain a very explicit analysis for  $d = 7$ , chosen as a relatively easy case, and then explain an efficient approach for arbitrary  $d$ .

In  $d = 7$ , the  $T$ -duality group is  $SO(3, 3)$ , which is the same as  $SL(4)$ ;  $U$ -duality extends this to  $G = SL(5)$ . A maximal compact subgroup is  $K = SO(5)$ .

In the NS-NS sector, there are six  $U(1)$  gauge fields that come from the compactification on a three-torus; they transform as a vector of  $SO(3, 3)$  or second rank antisymmetric tensor of  $SL(4)$ . In addition; four more  $U(1)$ ’s, transforming as a spinor of  $SO(3, 3)$  or a **4** of  $SL(4)$ , come from the RR sector. These states combine with the six from the NS-NS sector to make the second rank antisymmetric tensor, the **10** of  $SL(5)$ .

In Type II supergravity in seven dimensions, the maximal possible  $R$ -symmetry is  $K = SO(5)$  or  $Sp(4)$ . The supercharges make up in fact four pseudo-real spinors  $Q_\alpha^i$ ,  $i = 1 \dots 4$ , of the seven-dimensional Lorentz group  $SO(1, 6)$ , transforming as the **4** of

$Sp(4)$ . The central charges transform in the symmetric part of  $\mathbf{4} \times \mathbf{4}$ , which is the **10** or antisymmetric tensor of  $SO(5)$ . Thus, we are in a situation similar to what was described earlier in four dimensions: the gauge charges transform as the **10** of  $SL(5)$ , the central charges transform in the **10** of  $SO(5)$ , and a choice of vacuum in  $G/K = SL(5)/SO(5)$  selects an  $SO(5)$  subgroup of  $SL(5)$ , enabling one to identify these representations and map gauge charges to central charges.

A maximal abelian subgroup  $A$  of  $SL(5)$  is given by the diagonal matrices. A one-parameter subgroup of  $A$  consists of matrices of the form

$$g_t = \begin{pmatrix} e^{a_1 t} & 0 & 0 & 0 & 0 \\ 0 & e^{a_2 t} & 0 & 0 & 0 \\ 0 & 0 & e^{a_3 t} & 0 & 0 \\ 0 & 0 & 0 & e^{a_4 t} & 0 \\ 0 & 0 & 0 & 0 & e^{a_5 t} \end{pmatrix} \quad (3.11)$$

where the  $a_i$  are constants, not all zero, with  $\sum_i a_i = 0$ . We want to consider the behavior of the spectrum as  $t \rightarrow +\infty$ . By a Weyl transformation, we can limit ourselves to the case that

$$a_1 \geq a_2 \geq \dots \geq a_5. \quad (3.12)$$

Let  $\psi_{ij}$ ,  $i < j$  be a vector in the **10** of  $SL(5)$  whose components are zero except for the  $ij$  component, which is 1 (and the  $ji$  component, which is  $-1$ ). We will also use the name  $\psi_{ij}$  for a particle with those gauge charges. The mass formula  $M(\psi) = |Tg^{-1}\psi|$  says that the mass of  $\psi_{ij}$  scales with  $t$  as

$$M(\psi_{ij}) \sim e^{-t(a_i + a_j)}. \quad (3.13)$$

By virtue of (3.12), the lightest type of particle is  $\psi_{12}$ . For generic values of the  $a_i$ , this is the unique particle whose mass scales to zero fastest, but if  $a_2 = a_3$  then  $\psi_{12}$  is degenerate with other particles. To get a maximal set of particles degenerate with  $\psi_{12}$ , we need a maximal set of  $a_i$  equal to  $a_2$  and  $a_3$ . We cannot set all  $a_i$  equal (then they have to vanish, as  $\sum_i a_i = 0$ ), so by virtue of (3.12), there are two maximal cases, with  $a_1 = a_2 = a_3 = a_4$ , or  $a_2 = a_3 = a_4 = a_5$ . So the maximal degenerations correspond to

one-parameter subgroups

$$g_t = \begin{pmatrix} e^t & 0 & 0 & 0 & 0 \\ 0 & e^t & 0 & 0 & 0 \\ 0 & 0 & e^t & 0 & 0 \\ 0 & 0 & 0 & e^t & 0 \\ 0 & 0 & 0 & 0 & e^{-4t} \end{pmatrix} \quad (3.14)$$

or

$$g_t = \begin{pmatrix} e^{4t} & 0 & 0 & 0 & 0 \\ 0 & e^{-t} & 0 & 0 & 0 \\ 0 & 0 & e^{-t} & 0 & 0 \\ 0 & 0 & 0 & e^{-t} & 0 \\ 0 & 0 & 0 & 0 & e^{-t} \end{pmatrix} \quad (3.15)$$

with  $t \rightarrow +\infty$ . As we will see, the first corresponds to weakly coupled string theory, and the second to eleven-dimensional supergravity.

In (3.14), the particles whose masses vanish for  $t \rightarrow +\infty$  are the  $\psi_{ij}$  with  $1 \leq i < j \leq 4$ . There are six of these, the correct number of light elementary string states of string theory compactified from ten to seven dimensions. Moreover, in (3.14),  $g_t$  commutes with a copy of  $SL(4)$  that acts on indices  $1 - 2 - 3 - 4$ . This part of the seven-dimensional symmetry group  $SL(5)$  is unbroken by going to infinity in the direction (3.14), and hence would be observed as a symmetry of the low energy physics at “infinity” (though most of the symmetry is spontaneously broken in any given vacuum near infinity). Indeed,  $SL(4)$  with six gauge charges in the antisymmetric tensor representation is the correct  $T$ -duality group of weakly coupled string theory in seven dimensions.

There is a point here that may be puzzling at first sight. The full subgroup of  $SL(5)$  that commutes with  $g_t$  is actually not  $SL(4)$  but  $SL(4) \times \mathbf{R}^*$ , where  $\mathbf{R}^*$  is the one-parameter subgroup containing  $g_t$ . What happens to the  $\mathbf{R}^*$ ? When one restricts to the *integral* points in  $SL(5)$ , which are the true string symmetries, this  $\mathbf{R}^*$  does not contribute, so the symmetry group at infinity is just the integral form of  $SL(4)$ . A similar comment applies at several points below and will not be repeated.

Moving on now to the second case, in (3.15), the particles whose masses vanish for  $t \rightarrow +\infty$  are the  $\psi_{1i}$ ,  $i > 1$ . There are four of these, the correct number for compactification of eleven-dimensional supergravity on a four-torus  $\mathbf{T}^4$  whose dimensions are growing with  $t$ . The gauge charges of light states are simply the components of the momentum along  $\mathbf{T}^4$ .

The symmetry group at infinity is again  $SL(4)$ . This  $SL(4)$  has a natural interpretation as a group of linear automorphisms of  $\mathbf{T}^4$ .<sup>13</sup> In fact, the gauge charges carried by the light states in (3.15) transform in the **4** of  $SL(4)$ , which agrees with the supergravity description as that is how the momentum components along  $\mathbf{T}^4$  transform under  $SL(4)$ . As this  $SL(4)$  mixes three of the “original” ten dimensions with the eleventh dimension that is associated with strong coupling, we have our first evidence for the underlying eleven-dimensional Lorentz invariance.

Finally, let us consider a few non-maximal degenerations, to make sure we understand how to interpret them.<sup>14</sup> Degeneration in the direction

$$g_t = \begin{pmatrix} e^{3t} & 0 & 0 & 0 & 0 \\ 0 & e^{3t} & 0 & 0 & 0 \\ 0 & 0 & e^{-2t} & 0 & 0 \\ 0 & 0 & 0 & e^{-2t} & 0 \\ 0 & 0 & 0 & 0 & e^{-2t} \end{pmatrix} \quad (3.16)$$

leaves as  $t \rightarrow \infty$  the unique lightest state  $\psi_{12}$ . I interpret this as coming from partial decompactification to eight dimensions – taking one circle much larger than the others so that the elementary string states with momentum in that one direction are the lightest. This family has the symmetry group  $SL(3) \times SL(2)$ , which is indeed the  $U$ -duality group in eight dimensions, as it should be.

The family

$$g_t = \begin{pmatrix} e^{2t} & 0 & 0 & 0 & 0 \\ 0 & e^{2t} & 0 & 0 & 0 \\ 0 & 0 & e^{2t} & 0 & 0 \\ 0 & 0 & 0 & e^{-3t} & 0 \\ 0 & 0 & 0 & 0 & e^{-3t} \end{pmatrix} \quad (3.17)$$

gives three massless states  $\psi_{ij}$ ,  $1 \leq i < j \leq 3$ , transforming as **(3, 1)** of the symmetry group  $SL(3) \times SL(2)$ . I interpret this as decompactification to the Type IIB theory in

<sup>13</sup> That is, if  $\mathbf{T}^4$  is understood as the space of real variables  $y^i$ ,  $i = 1 \dots 4$ , modulo  $y^i \rightarrow y^i + n^i$ , with  $n^i \in \mathbf{Z}$ , then  $SL(4)$  acts by  $y^i \rightarrow w^i_j y^j$ . For this to be a diffeomorphism and preserve the orientation, the determinant of  $w$  must be one, so one is in  $SL(4)$ . Given an  $n$ -torus  $\mathbf{T}^n$ , we will subsequently use the phrase “mapping class group” to refer to the  $SL(n)$  that acts linearly in this sense on  $\mathbf{T}^n$ .

<sup>14</sup> We will see in the next section that when the  $U$ -duality group has rank  $r$ , there are  $r$  naturally distinguished one-parameter subgroups. For  $SL(5)$ , these are (3.14), (3.15), and the two introduced below.

ten dimensions – taking all three circles to be very large. The three light charges are the momenta around the three circles;  $SL(3)$  is the mapping class group of the large three-torus, and  $SL(2)$  is the  $U$ -duality group of the Type IIB theory in ten dimensions.

### *Partially Saturated States*

I will now justify an assumption made above and also make a further test of the interpretation that we have proposed.

First of all, we identified BPS-saturated elementary string states with charge tensors  $\psi_{ij}$  with (in the right basis) only one non-zero entry. Why was this valid?

We may as well consider NS-NS states; then we can restrict ourselves to the  $T$ -duality group  $SO(3, 3)$ . The gauge charges transform in the vector representation of  $SO(3, 3)$ . Given such a vector  $v_a$ , one can define the quadratic invariant  $(v, v) = \sum_{a,b} \eta^{ab} v_a v_b$ .

On the other hand,  $SO(3, 3)$  is the same as  $SL(4)$ , and  $v$  is equivalent to a second rank antisymmetric tensor  $\psi$  of  $SL(4)$ . In terms of  $\psi$ , the quadratic invariant is  $(\psi, \psi) = \frac{1}{4} \epsilon^{ijkl} \psi_{ij} \psi_{kl}$ . By an  $SL(4)$  transformation, one can bring  $\psi$  to a normal form in which the independent non-zero entries are  $\psi_{12}$  and  $\psi_{34}$  only. Then

$$(\psi, \psi) = 2\psi_{12}\psi_{34}. \quad (3.18)$$

So the condition that the particle carries only one type of charge, that is, that only  $\psi_{12}$  or  $\psi_{34}$  is non-zero, is that  $(\psi, \psi) = 0$ .

Now let us consider the elementary string states. Such a state has in the toroidal directions left- and right-moving momenta  $p_L$  and  $p_R$ .  $p_L$  and  $p_R$  together form a vector of  $SO(3, 3)$ , and the quadratic invariant is [8]

$$(p, p) = |p_L|^2 - |p_R|^2. \quad (3.19)$$

BPS-saturated states have no oscillator excitations for left- or right-movers, and the mass shell condition requires that they obey  $|p_L|^2 - |p_R|^2 = 0$ , that is, that the momentum or charge vector  $p$  is light-like. This implies, according to the discussion in the last paragraph, that in the right basis, the charge tensor  $\psi$  has only one entry. That is the assumption we made.

Now, however, we can do somewhat better and consider elementary string states of Type II that are BPS-saturated for left-movers only (or equivalently, for right-movers only). Such states are in “middle-sized” supermultiplets, of dimension  $2^{12}$  (as opposed to generic supermultiplets of dimension  $2^{16}$  and BPS-saturated multiplets of dimension  $2^8$ ). To achieve BPS saturation for the left-movers only, one puts the left-moving oscillators in their ground state, but one permits right-moving oscillator excitations; as those excitations are arbitrary, one gets an exponential spectrum of these half-saturated states (analogous to the exponential spectrum of BPS-saturated states in the heterotic string [26]). With oscillator excitations for right-movers only, the mass shell condition implies that  $|p_L|^2 > |p_R|^2$ , and hence the charge vector is not lightlike. The charge tensor  $\psi$  therefore in its normal form has both  $\psi_{12}$  and  $\psi_{34}$  non-zero. For such states, the mass inequality says that the mass is bounded below by the largest eigenvalue of  $Tg^{-1}\psi$ , with equality for the “middle-sized” multiplets.

With this in mind, let us consider the behavior of such half-saturated states in the various degenerations. In the “stringy” degeneration (3.14), a state with non-zero  $\psi_{12}$  and  $\psi_{34}$  has a mass of the same order of magnitude as a state with only  $\psi_{12}$  non-zero. This is as we would expect from weakly coupled string theory with toroidal radii of order one: the half-saturated states have masses of the same order of magnitude as the BPS-saturated massive modes. To this extent, string excitations show up in the strong coupling analysis.

What about the “eleven-dimensional” degeneration (3.15)? In this case, while the particles with only one type of charge have masses that vanish as  $e^{-3t}$  for  $t \rightarrow \infty$ , the particles with two kinds of charge have masses that grow as  $e^{+t}$ . The only light states that we can see with this formalism in this degeneration are the Kaluza-Klein modes of eleven-dimensional supergravity. There is, for instance, no evidence for membrane excitations; such evidence might well have appeared if a consistent membrane theory with eleven-dimensional supergravity as its low energy limit really does exist.

### *3.5. Framework For General Analysis*

It would be tiresome to repeat this analysis “by hand” in other values of the dimension.

Instead, I will now<sup>15</sup> explain a bit of group theory that makes the analysis easy. One of the main points is to incorporate the action of the Weyl group. This was done above by choosing  $a_1 \geq a_2 \geq \dots \geq a_5$ , but to exploit the analogous condition in  $E_7$ , for instance, a little machinery is useful.

In  $d$  dimensions, the  $U$ -duality group  $G$  has rank  $r = 11 - d$ . Given any one-parameter subgroup  $F$  of a maximal abelian subgroup  $A$ , one can pick a set of simple positive roots  $x_i$  such that the action of  $F$  on the  $x_i$  is

$$x_i \rightarrow e^{c_i t} x_i \quad (3.20)$$

with  $c_i$  non-negative. In this restriction on the  $c_i$ , we have used the Weyl action. Conversely, for every set of non-negative  $c_i$  (not all zero), there is a one-parameter subgroup  $F$  that acts as (3.20).

The gauge charges are in some representation  $R$  of  $G$ ; that is, for each weight in  $R$  there is a corresponding gauge charge.<sup>16</sup> Let  $\rho = \sum_i e_i x_i$  be the highest weight in  $R$ . The  $e_i$  are positive integers. A particle whose only gauge charge is the one that corresponds to  $\rho$  has a mass that vanishes for  $t \rightarrow +\infty$  as

$$M_\rho \sim \exp \left( - \sum_i c_i e_i t \right). \quad (3.21)$$

Any other weight in  $R$  is of the form  $\rho' = \sum_i f_i x_i$ , with  $f_i \leq e_i$ . A particle carrying the  $\rho'$  charge has mass of order

$$M_{\rho'} \sim \exp \left( - \sum_i c_i f_i t \right). \quad (3.22)$$

Thus  $M_{\rho'} \geq M_\rho$  – the particle with only charge  $\rho$  always goes to zero mass at least as fast as any other – and  $M_{\rho'} = M_\rho$  if and only if

$$c_i = 0 \text{ whenever } f_i < e_i. \quad (3.23)$$

<sup>15</sup> With some assistance from A. Borel.

<sup>16</sup> The particular representations  $R$  that actually arise in Type II string theory in  $d \geq 4$  have the property (unusual among representations of Lie groups) that the non-zero weight spaces are all one-dimensional. It therefore makes sense to label the gauge charges by weights. (These representations are actually “minuscule” – the Weyl group acts transitively on the weights.)  $d \leq 3$  would have some new features, as already mentioned above.

Now, our problem is to pick the subgroup  $F$ , that is, the  $c_i$ , so that a maximal set of  $M_{\rho'}$  are equal to  $M_\rho$ . If the  $c_i$  are all non-zero, then (as the highest weight state is unique) (3.23) implies that  $\rho' = \rho$  and only one gauge charge is carried by the lightest particles. The condition in (3.23) becomes less restrictive only when one of the  $c_i$  becomes zero, and to get a maximal set of  $M_{\rho'}$  degenerate with  $M_\rho$ , we must set as many of the  $c_i$  as possible to zero. As the  $c_i$  may not all vanish, the best we can do is to set  $r - 1$  of them to zero. There are therefore precisely  $r$  one-parameter subgroups  $F_i$  to consider, labeled by which of the  $c_i$  is non-zero.

The  $x_i$  are labeled by the vertices in the Dynkin diagram of  $G$ , so each of the  $F_i$  is associated with a particular vertex  $P_i$ . Deleting  $P_i$  from the Dynkin diagram of  $G$  leaves the Dynkin diagram of a rank  $r - 1$  subgroup  $H_i$  of  $G$ . It is the unbroken subgroup when one goes to infinity in the  $F_i$  direction.

### 3.6. Analysis In $d = 4$

With this machinery, it is straightforward to analyze the dynamics in each dimension  $d$ . As the rank is  $r = 11 - d$ , there are  $11 - d$  distinguished one-parameter subgroups to check. It turns out that one of them corresponds to weakly coupled string theory in  $d$  dimensions, one to toroidal compactification of eleven-dimensional supergravity to  $d$  dimensions, and the others to partial (or complete) decompactifications. In each case, the symmetry group when one goes to infinity is the expected one: the  $T$ -duality group  $SO(10 - d, 10 - d)$  for the string degeneration; the mapping class group  $SL(11 - d)$  for supergravity; or for partial decompactification to  $d'$  dimensions, the product of the mapping class group  $SL(d' - d)$  of a  $d' - d$ -torus and the  $U$ -duality group in  $d'$  dimensions.

I will illustrate all this in  $d = 4$ , where the  $U$ -duality group is  $E_7$ . Going to infinity in a direction  $F_i$  associated with one of the seven points in the Dynkin diagram leaves as unbroken subgroup  $H_i$  one of the following:

- (1)  $SO(6, 6)$ : this is the  $T$ -duality group for string theory toroidally compactified from ten to four dimensions. This is a maximal degeneration, with (as we will see) 12 massless states transforming in the **12** of  $SO(6, 6)$ .
- (2)  $SL(7)$ : this is associated with eleven-dimensional supergravity compactified to

four dimensions on a seven-torus whose mapping class group is  $SL(7)$ . This is the other maximal degeneration; there are the expected seven massless states in the **7** of  $SL(7)$ .

(3)  $E_6$ : this and the other cases are non-maximal degenerations corresponding to partial decompactification. This case corresponds to partial decompactification to five dimensions by taking one circle to be much larger than the others; there is only one massless state, corresponding to a state with momentum around the large circle.  $E_6$  arises as the  $U$ -duality group in five dimensions.

(4)  $SL_2 \times SO(5, 5)$ : this is associated with partial decompactification to six dimensions. There are two light states, corresponding to momenta around the two large circles; they transform as **(2, 1)** under  $SL_2 \times SO(5, 5)$ .  $SL_2$  acts on the two large circles and  $SO(5, 5)$  is the  $U$ -duality group in six dimensions.

(5)  $SL_3 \times SL(5)$ : this is associated with partial decompactification to seven dimensions.  $SL(3)$  acts on the three large circles (and the three light charges), and  $SL(5)$  is the  $U$ -duality in seven dimensions.

(6)  $SL_4 \times SL(3) \times SL(2)$ : this is associated with partial decompactification to eight dimensions.  $SL(4)$  acts on the four large circles and light charges, and  $SL(3) \times SL(2)$  is  $U$ -duality in eight dimensions.

(7)  $SL_6 \times SL_2$ : this is associated with decompactification to Type IIB in ten dimensions.  $SL_6$  acts on the six large circles and light charges, and  $SL(2)$  is the  $U$ -duality in ten dimensions.

In what follows, I will just check the assertions about the light spectrum for the first two cases, which are the important ones, and the third, which is representative of the others.

(1)  $F_1$  can be described as follows.  $E_7$  contains a maximal subgroup  $SO(6, 6) \times SL(2)$ .  $F_1$  can be taken as the subgroup of  $SL(2)$  consisting of matrices of the form

$$\begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix}. \quad (3.24)$$

The gauge charges are in the **56** of  $E_7$ , which decomposes under  $L_1$  as **(12, 2)**  $\oplus$  **(32, 1)**. The lightest states come from the part of the **(12, 2)** that transforms as  $e^t$  under (3.24); these are the expected twelve states in the **12** of  $SO(6, 6)$ .

(2)  $E_7$  contains a maximal subgroup  $SL(8)$ .  $F_2$  can be taken as the subgroup of  $SL(8)$  consisting of group elements  $g_t = \text{diag}(e^t, e^t, \dots, e^t, e^{-7t})$ . The **56** of  $E_7$  decomposes as **28** $\oplus$ **28'** – the antisymmetric tensor plus its dual. The states of highest eigenvalue (namely  $e^{8t}$ ) are seven states in the **28** transforming in the expected **7** of the unbroken  $SL(7)$ .

(3)  $E_7$  has a maximal subgroup  $E_6 \times \mathbf{R}^*$ , and  $F_3$  is just the  $\mathbf{R}^*$ . The **56** of  $E_7$  decomposes as **27** $^1\oplus$ **27'** $^{-1}\oplus$ **1** $^3\oplus$ **1** $^{-3}$ , where the  $E_6$  representation is shown in boldface and the  $\mathbf{R}^*$  charge (with some normalization) by the exponent. Thus in the  $F_3$  degeneration, there is a unique lightest state, the **1** $^3$ .

The reader can similarly analyze the light spectrum for the other  $F_i$ , or the analogous subgroups in  $d \neq 4$ .

## 4. Heterotic String Dynamics Above Four Dimensions

### 4.1. A Puzzle In Five Dimensions

$S$ -duality gives an attractive proposal for the strong coupling dynamics of the heterotic string after toroidal compactification to four dimensions: it is equivalent to the same theory at weak coupling. In the remainder of this paper, we will try to guess the behavior above four dimensions. This process will also yield some new insight about  $S$ -duality in four dimensions.

Toroidal compactification of the heterotic string from 10 to  $d$  dimensions gives  $2(10-d)$  vectors that arise from dimensional reduction of the metric and antisymmetric tensor. Some of the elementary string states are electrically charged with respect to these vectors.

Precisely in five dimensions, one more vector arises. This is so because in five dimensions a two-form  $B_{mn}$  is dual to a vector  $A_m$ , roughly by  $dB = *dA$ . In the elementary string spectrum, there are no particles that are electrically charged with respect to  $A$ , roughly because  $A$  can be defined (as a vector) only in five dimensions. But it is easy to see where to find such electric charges. Letting  $H$  be the field strength of  $B$  (including the Chern-Simons terms) the anomaly equation

$$dH = \text{tr}F \wedge F - \text{tr}R \wedge R \quad (4.1)$$

( $F$  is the  $E_8 \times E_8$  or  $SO(32)$  field strength and  $R$  the Riemann tensor) implies that the electric current of  $A$  is

$$J = * \text{tr} F \wedge F - * \text{tr} R \wedge R. \quad (4.2)$$

Thus, with  $G = dA$ , (4.1) becomes

$$D^m G_{mn} = J_n, \quad (4.3)$$

showing that  $J_n$  is the electric current. So the charge density  $J_0$  is the instanton density, and a Yang-Mills instanton, regarded as a soliton in 4+1 dimensions, is electrically charged with respect to  $A$ .

Instantons (and their generalizations to include the supergravity multiplet [27,28]) are invariant under one half of the supersymmetries. One would therefore suspect that quantization of the instanton would give BPS-saturated multiplets, with masses given by the instanton action:

$$M = \frac{16\pi^2|n|}{\lambda^2}. \quad (4.4)$$

Here  $n$  is the instanton number or electric charge and  $\lambda$  is the string coupling.

To really prove existence of these multiplets, one would need to understand and quantize the collective coordinates of the stringy instanton. In doing this, one needs to pick a particular vacuum to work in. In the generic toroidal vacuum, the unbroken gauge group is just a product of  $U(1)$ 's. Then the instantons, which require a non-abelian structure, tend to shrink to zero size, where stringy effects are strong and the analysis is difficult. Alternatively, one can consider a special vacuum with an unbroken non-abelian group, but this merely adds infrared problems to the stringy problems. The situation is analogous to the study [29] of  $H$ -monopoles after toroidal compactification to four dimensions; indeed, the present paper originated with an effort to resolve the problems concerning  $H$ -monopoles. (The connection between instantons and  $H$ -monopoles is simply that upon compactification of one of the spatial directions on a circle, the instantons become what have been called  $H$ -monopoles.)

Despite the difficulty in the collective coordinate analysis, there are two good reasons to believe that BPS-saturated multiplets in this sector do exist. One, already mentioned,

is the invariance of the classical solution under half the supersymmetries. The second reason is that if in five dimensions, the electrically charged states had masses bounded strictly above the BPS value in (4.4), the same would be true after compactification on a sufficiently big circle, and then the BPS-saturated  $H$ -monopoles required for  $S$ -duality could not exist.

Accepting this assumption, we are in a similar situation to that encountered earlier for the Type IIA string in ten dimensions: there is a massless vector, which couples to electric charges whose mass diverges for weak coupling. (The mass is here proportional to  $1/\lambda^2$  in contrast to  $1/\lambda$  in the other case.) Just as in the previous situation, we have a severe puzzle if we take the formula seriously for strong coupling, when these particles seem to go to zero mass.

If we are willing to take (4.4) seriously for strong coupling, then we have for each integer  $n$  a supermultiplet of states of charge  $n$  and mass proportional to  $|n|$ , going to zero mass as  $\lambda \rightarrow \infty$ . It is very hard to interpret such a spectrum in terms of local field theory in five dimensions. But from our previous experience, we know what to do: interpret these states as Kaluza-Klein states on  $\mathbf{R}^5 \times \mathbf{S}^1$ .

The  $\mathbf{S}^1$  here will have to be a “new” circle, not to be confused with the five-torus  $\mathbf{T}^5$  in the original toroidal compactification to five dimensions. (For instance, the  $T$ -duality group  $SO(21, 5)$  acts on  $\mathbf{T}^5$  but not on the new circle.) So altogether, we seem to have eleven dimensions,  $\mathbf{R}^5 \times \mathbf{S}^1 \times \mathbf{T}^5$ , and hence we seem to be in need of an eleven-dimensional supersymmetric theory.

In section two, eleven-dimensional supergravity made a handy appearance at this stage, but here we seem to be in a quandary. There is no obvious way to introduce an eleventh dimension relevant to the heterotic string. Have we reached a dead end?

#### 4.2. The Heterotic String In Six Dimensions

Luckily, there is a conjectured relation between the heterotic string and Type II superstrings [3,4] which has just the right properties to solve our problem (though not by leading us immediately back to eleven dimensions). The conjecture is that the heterotic string toroidally compactified to *six* dimensions is equivalent to the Type IIA superstring

compactified to six dimensions on a K3 surface.

The evidence for this conjecture has been that both models have the same supersymmetry and low energy spectrum in six dimensions and the same moduli space of vacua, namely  $SO(20, 4; \mathbf{Z}) \backslash SO(20, 4; \mathbf{R}) / (SO(20) \times SO(4))$ . For the toroidally compactified heterotic string, this structure for the moduli space of vacua is due to Narain [8]; for Type II, the structure was determined locally by Seiberg [30] and globally by Aspinwall and Morrison [31].

In what follows, I will give several new arguments for this “string-string duality” between the heterotic string and Type IIA superstrings:

(1) When one examines more precisely how the low energy effective actions match up, one finds that weak coupling of one theory corresponds to strong coupling of the other theory. This is a necessary condition for the duality to make sense, since we certainly know that the heterotic string for weak coupling is not equivalent to the Type IIA superstring for weak coupling.

(2) Assuming string-string duality in six dimensions, we will be able to resolve the puzzle about the strong coupling dynamics of the heterotic string in five dimensions. The strongly coupled heterotic string on  $\mathbf{R}^5$  (times a five-torus whose parameters are kept fixed) is equivalent to a Type IIB superstring on  $\mathbf{R}^5 \times \mathbf{S}^1$  (times a K3 whose parameters are kept fixed). The effective six-dimensional Type IIB theory is weakly coupled at its compactification scale, so this is an effective solution of the problem of strong coupling for the heterotic string in five dimensions.

(3) We will also see that – as anticipated by Duff in a more abstract discussion [7] – string-string duality in six dimensions implies  $S$ -duality of the heterotic string in four dimensions. Thus, all evidence for  $S$ -duality can be interpreted as evidence for string-string duality, and one gets at least a six-dimensional answer to the question “what higher dimensional statement leads to  $S$ -duality in four dimensions?”

(4) The K3 becomes singular whenever the heterotic string gets an enhanced symmetry group; the singularities have an  $A - D - E$  classification, just like the enhanced symmetries.

(5) Finally, six-dimensional string-string duality also leads to an attractive picture for heterotic string dynamics in seven dimensions. (Above seven dimensions the analysis

would be more complicated.)

I would like to stress that some of these arguments test more than a long distance relation between the heterotic string and strongly coupled Type IIA. For instance, in working out the five-dimensional dynamics via string-string duality, we will be led to a Type IIA theory with a *small* length scale, and to get a semi-classical description will require a  $T$ -duality transformation, leading to Type IIB. The validity of the discussion requires that six-dimensional string-string duality should be an exact equivalence, like the  $SL(2, \mathbf{Z})$  symmetry for Type IIB in ten dimensions and unlike the relation of Type II to eleven-dimensional supergravity.

#### 4.3. Low Energy Actions

Let us start by writing a few terms in the low energy effective action of the heterotic string, toroidally compactified to six dimensions. We consider the metric  $g$ , dilaton  $\phi$ , and antisymmetric tensor field  $B$ , and we let  $C$  denote a generic abelian gauge field arising from the toroidal compactification. We are only interested in keeping track of how the various terms scale with  $\phi$ . For the heterotic string, the whole classical action scales as  $e^{-2\phi} \sim \lambda^{-2}$ , so one has very roughly

$$I = \int d^6x \sqrt{g} e^{-2\phi} (R + |\nabla\phi|^2 + |dB|^2 + |dC|^2). \quad (4.5)$$

On the other hand, consider the Type IIA superstring in six dimensions. The low energy particle content is the same as for the toroidally compactified heterotic string, at least at a generic point in the moduli space of the latter where the unbroken gauge group is abelian. Everything is determined by  $N = 4$  supersymmetry except the number of  $U(1)$ 's in the gauge group and the number of antisymmetric tensor fields; requiring that these match with the heterotic string leads one to use Type IIA rather than Type IIB. So in particular, the low energy theory derived from Type IIA has a dilaton  $\phi'$ , a metric  $g'$ , an antisymmetric tensor field  $B'$ , and gauge fields  $C'$ . <sup>17</sup> Here  $\phi'$ ,  $g'$ , and  $B'$  come from the

---

<sup>17</sup> We normalize  $B'$  and  $C'$  to have standard gauge transformation laws. Their gauge transformations would look different if one scaled the fields by powers of  $e^\phi$ . This point was discussed in section two.

NS-NS sector, but  $C'$  comes from the RR sector, so as we noted in section two, the kinetic energy of  $\phi'$ ,  $g'$ , and  $B'$  scales with the dilaton just like that in (4.5), but the kinetic energy of  $C'$  has no coupling to the dilaton. So we have schematically

$$I' = \int d^6x \sqrt{g'} \left( e^{-2\phi'} (R' + |\nabla\phi'|^2 + |dB'|^2) + |dC'|^2 \right). \quad (4.6)$$

We need the change of variables that turns (4.5) into (4.6). In (4.5), the same power of  $e^\phi$  multiplies  $R$  and  $|dC'|^2$ . We can achieve that result in (4.6) by the change of variables  $g' = g''e^{2\phi'}$ . Then (4.6) becomes

$$I' = \int d^6x \sqrt{g''} \left( e^{2\phi'} (R'' + |\nabla\phi'|^2) + e^{-2\phi'} |dB'|^2 + e^{2\phi'} |dC'|^2 \right). \quad (4.7)$$

Now the coefficient of the kinetic energy of  $B'$  is the opposite of what we want, but this can be reversed by a duality transformation. The field equations of  $B'$  say that  $d*(e^{-2\phi'} dB') = 0$ , so the duality transformation is

$$e^{-2\phi'} dB' = *dB''. \quad (4.8)$$

Then (4.7) becomes

$$I' = \int d^6x \sqrt{g''} e^{2\phi'} (R'' + |\nabla\phi'|^2 + |dB''|^2 + |dC'|^2). \quad (4.9)$$

This agrees with (4.5) if we identify  $\phi = -\phi'$ . Putting everything together, the change of variables by which one can identify the low energy limits of the two theories is

$$\begin{aligned} \phi &= -\phi' \\ g &= e^{2\phi} g' = e^{-2\phi'} g' \\ dB &= e^{-2\phi'} * dB' \\ C &= C'. \end{aligned} \quad (4.10)$$

Unprimed and primed variables are fields of the heterotic string and Type IIA, respectively.

In particular, the first equation implies that weak coupling of one theory is equivalent to strong coupling of the other. This makes it possible for the two theories to be equivalent without the equivalence being obvious in perturbation theory.

#### 4.4. Dynamics In Five Dimensions

Having such a (conjectured) exact statement in six dimensions, one can try to deduce the dynamics below six dimensions. The ability to do this is not automatic because (just as in field theory) the dimensional reduction might lead to new dynamical problems at long distances. But we will see that in this particular case, the string-string duality in six dimensions does determine what happens in five and four dimensions.

We first compactify the heterotic string from ten to six dimensions on a torus (which will be kept fixed and not explicitly mentioned), and then take the six-dimensional world to be  $\mathbf{R}^5 \times \mathbf{S}_r^1$ , where  $\mathbf{S}_r^1$  will denote a circle of radius  $r$ . We want to keep  $r$  fixed and take  $\lambda = e^\phi$  to infinity. According to (4.10), the theory in this limit is equivalent to the Type IIA superstring on  $\mathbf{R}^5 \times \mathbf{S}_r^1$ , times a K3 surface (of fixed moduli), with string coupling and radius  $\lambda'$  and  $r'$  given by

$$\begin{aligned}\lambda' &= \lambda^{-1} \\ r' &= \lambda^{-1}r.\end{aligned}\tag{4.11}$$

In particular, the coupling  $\lambda'$  goes to zero in the limit for  $\lambda \rightarrow \infty$ . However, the radius  $r'$  in the dual theory is also going to zero. The physical interpretation is much clearer if one makes a  $T$ -duality transformation, replacing  $r'$  by

$$r'' = \frac{1}{r'} = \frac{\lambda}{r}.\tag{4.12}$$

The  $T$ -duality transformation also acts on the string coupling constant. This can be worked out most easily by noting that the effective five-dimensional gravitational constant, which is  $\lambda^2/r$ , must be invariant under the  $T$ -duality. So under  $r' \rightarrow 1/r'$ , the string coupling  $\lambda'$  is replaced by

$$\lambda'' = \frac{\lambda'}{r'}\tag{4.13}$$

so that

$$\frac{r'}{(\lambda')^2} = \frac{r''}{(\lambda'')^2}.\tag{4.14}$$

Combining this with (4.11), we learn that the heterotic string on  $\mathbf{R}^5 \times \mathbf{S}_r^1$  and string coupling  $\lambda$  is equivalent to a Type II superstring with coupling and radius

$$\begin{aligned}\lambda'' &= r^{-1} \\ r'' &= \frac{\lambda}{r}.\end{aligned}\tag{4.15}$$

This is actually a Type IIB superstring, since the  $T$ -duality transformation turns the Type IIA model that appears in the string-string duality conjecture in six dimensions into a Type IIB superstring.

(4.15) shows that the string coupling constant of the effective Type IIB theory remains fixed as  $\lambda \rightarrow \infty$  with fixed  $r$ , so the dual theory is not weakly coupled at all length scales. However, (4.15) also shows that  $r'' \rightarrow \infty$  in this limit, and this means that at the length scale of the compactification, the effective coupling is weak. (The situation is similar to the discussion of the strongly coupled ten-dimensional Type IIA superstring in section two.) All we need to assume is that the six-dimensional Type II superstring theory, even with a coupling of order one, is equivalent at long distances to weakly coupled Type II supergravity. If that is so, then when compactified on a very large circle, it can be described at and above the compactification length by the weakly coupled supergravity, which describes the dynamics of the light degrees of freedom.

### *Moduli Space Of Type IIB Vacua*

The following remarks will aim to give a more fundamental explanation of (4.15) and a further check on the discussion.

Consider the compactification of Type IIB superstring theory on  $\mathbf{R}^6 \times K3$ . This gives a chiral  $N = 4$  supergravity theory in six dimensions, with five self-dual two-forms (that is, two-forms with self-dual field strength) and twenty-one anti-self-dual two-forms (that is, two-forms with anti-self-dual field strength). The moduli space of vacua of the low energy supergravity theory is therefore [32]  $G/K$  with  $G = SO(21, 5)$  and  $K$  the maximal subgroup  $SO(21) \times SO(5)$ .

The coset space  $G/K$  has dimension  $21 \times 5 = 105$ . The interpretation of this number is as follows. There are 80 NS-NS moduli in the conformal field theory on  $K3$  (that, the moduli space of  $(4, 4)$  conformal field theories on  $K3$  is 80-dimensional). There are 24 zero modes of RR fields on  $K3$ . Finally, the expectation value of the dilaton – the string coupling constant – gives one more modulus. In all, one has  $80 + 24 + 1 = 105$  states. In particular, the string coupling constant is unified with the others.

It would be in the spirit of  $U$ -duality to suppose that the Type IIB theory on  $\mathbf{R}^6 \times K3$

has the discrete symmetry group  $SO(21, 5; \mathbf{Z})$ . In fact, that follows from the assumption of  $SL(2, \mathbf{Z})$  symmetry of Type IIB in ten dimensions [3] together with the demonstration in [31] of a discrete symmetry  $SO(20, 4; \mathbf{Z})$  for  $(4, 4)$  conformal field theories on K3. For the  $SO(20, 4; \mathbf{Z})$  and  $SL(2, \mathbf{Z})$  do not commute and together generate  $SO(21, 5; \mathbf{Z})$ . The moduli space of Type IIB vacua on  $\mathbf{R}^6 \times K3$  is hence

$$\mathcal{N} = SO(21, 5; \mathbf{Z}) \backslash SO(21, 5; \mathbf{R}) / (SO(21) \times SO(5)). \quad (4.16)$$

Now consider the Type IIB theory on  $\mathbf{R}^5 \times \mathbf{S}^1 \times K3$ . One gets one new modulus from the radius of the  $\mathbf{S}^1$ . No other new moduli appear (the Type IIB theory on  $\mathbf{R}^6 \times K3$  has no gauge fields so one does not get additional moduli from Wilson lines). So the moduli space of Type IIB vacua on  $\mathbf{R}^5 \times \mathbf{S}^1 \times K3$  is

$$\mathcal{M} = \mathcal{N} \times \mathbf{R}^+, \quad (4.17)$$

where  $\mathbf{R}^+$  (the space of positive real numbers) parametrizes the radius of the circle.

What about the heterotic string on  $\mathbf{R}^5 \times \mathbf{T}^5$ ? The  $T$ -duality moduli space of the toroidal vacua is precisely  $\mathcal{N} = SO(21, 5; \mathbf{Z}) \backslash SO(21, 5; \mathbf{R}) / (SO(21) \times SO(5))$ . There is one more modulus, the string coupling constant. So the moduli space of heterotic string vacua on  $\mathbf{R}^5 \times \mathbf{T}^5$  is once again  $\mathcal{M} = \mathcal{N} \times \mathbf{R}^+$ . Now the  $\mathbf{R}^+$  parametrizes the string coupling constant.

So the moduli space of toroidal heterotic string vacua on  $\mathbf{R}^5 \times \mathbf{T}^5$  is the same as the moduli space of Type IIB vacua on  $\mathbf{R}^5 \times K3$ , suggesting that these theories may be equivalent. The map between them turns the string coupling constant of the heterotic string into the radius of the circle in the Type IIB description. This is the relation that we have seen in (4.15) (so, in particular, strong coupling of the heterotic string goes to large radius in Type IIB).

To summarize the discussion, we have seen that an attractive conjecture – the equivalence of the heterotic string in six dimensions to a certain Type IIA theory – implies another attractive conjecture – the equivalence of the heterotic string in five dimensions to a certain Type IIB theory. The link from one conjecture to the other depended on a

$T$ -duality transformation, giving evidence that these phenomena must be understood in terms of string theory, not just in terms of relations among low energy field theories.

### Detailed Matching Of States

Before leaving this subject, perhaps it would be helpful to be more explicit about how the heterotic and Type II spectra match up in five dimensions.

Compactification of the six-dimensional heterotic string theory on  $\mathbf{R}^5 \times \mathbf{S}^1$  generates in the effective five-dimensional theory three  $U(1)$  gauge fields that were not present in six dimensions. There is the component  $g_{m6}$  of the metric, the component  $B_{m6}$  of the antisymmetric tensor field, and the vector  $A_m$  that is dual to the spatial components  $B_{mn}$  of the antisymmetric tensor field. Each of these couples to charged states:  $g_{m6}$  couples to elementary string states with momentum around the circle,  $B_{m6}$  to states that wind around the circle, and  $A_m$  to states that arise as instantons in four spatial dimensions, invariant under rotations about the compactified circle. The mass of these “instantons” is  $r/\lambda^2$ , with the factor of  $r$  coming from integrating over the circle and  $1/\lambda^2$  the instanton action in four dimensions. The masses of these three classes of states are hence of order  $1/r$ ,  $r$ , and  $r/\lambda^2$ , respectively, if measured with respect to the string metric. To compare to Type II, we should remember (4.10) that a Weyl transformation  $g = \lambda^2 g'$  is made in going to the sigma model metric of the Type IIA description. This multiplies masses by a factor of  $\lambda$ , so the masses computed in the heterotic string theory but measured in the string units of Type IIA are

$$\begin{aligned} g_{m6} &: \frac{\lambda}{r} \\ B_{m6} &: \lambda r \\ A_m &: \frac{r}{\lambda}. \end{aligned} \tag{4.18}$$

Likewise, compactification of the six-dimensional Type IIA superstring on  $\mathbf{R}^5 \times \mathbf{S}^1$  gives rise to three vectors  $g'_{m6}$ ,  $B'_{m6}$ , and  $A'$ . The first two couple to elementary string states. The last presumably couples to some sort of soliton, perhaps the classical solution that has been called the symmetric five-brane [28]. Its mass would be of order  $r'/(λ')^2$  in string units for the same reasons as before. The masses of particles coupling to the three

vectors are thus in string units:

$$\begin{aligned} g'_{m6} &: \frac{1}{r'} \\ B'_{m6} &: r' \\ A'_m &: \frac{r'}{(\lambda')^2}. \end{aligned} \tag{4.19}$$

Now, (4.18) agrees with (4.19) under the expected transformation  $\lambda = 1/\lambda'$ ,  $r = \lambda r'$  provided that one identifies  $g_{m6}$  with  $g'_{m6}$ ;  $B_{m6}$  with  $A'_m$ ; and  $A_m$  with  $B'_{m6}$ . The interesting point is of course that  $B_{m6}$  and  $A_m$  switch places. But this was to be expected from the duality transformation  $dB \sim *dB'$  that enters in comparing the two theories.

So under string-string duality the “instanton,” which couples to  $A_m$ , is turned into the string winding state, which couples to  $B'_{m6}$ , and the string winding state that couples to  $B_{m6}$  is turned into a soliton that couples to  $A'_m$ .

#### 4.5. Relation To S-Duality

Now we would like to use six-dimensional string-string duality to determine the strong coupling dynamics of the heterotic string in four dimensions. Once again, we start with a preliminary toroidal compactification from ten to six dimensions on a fixed torus that will not be mentioned further. Then we take the six-dimensional space to be a product  $\mathbf{R}^4 \times \mathbf{T}^2$ , with  $\mathbf{T}^2$  a two-torus. String-string duality says that this is equivalent to a six-dimensional Type IIA theory on  $\mathbf{R}^4 \times \mathbf{T}^2$  (with four extra dimensions in the form of a fixed K3).

One might now hope, as in six and five dimensions, to take the strong coupling limit and get a useful description of strongly coupled four-dimensional heterotic string theory in terms of Type II. This fails for the following reason. In six dimensions, the duality related strong coupling of the heterotic string to weak coupling of Type IIA. In five dimensions, it related weak coupling of the heterotic string to coupling of order one of Type IIB (see (4.15)). Despite the coupling of order one, this was a useful description because the radius of the sixth dimension was large, so (very plausibly) the effective coupling at the compactification scale is small. A similar scaling in four dimensions, however, will show that the strong coupling limit of the heterotic string in four dimensions is related to a

strongly coupled four-dimensional Type II superstring theory, and now one has no idea what to expect.

It is remarkable, then, that there is another method to use six-dimensional string-string duality to determine the strong coupling behavior of the heterotic string in four dimensions. This was forecast and explained by Duff [7] without reference to any particular example. The reasoning goes as follows.

Recall (such matters are reviewed in [33]) that the  $T$ -duality group of a two-torus is  $SO(2, 2)$  which is essentially the same as  $SL(2) \times SL(2)$ . Here the two  $SL(2)$ 's are as follows. One of them, sometimes called  $SL(2)_U$ , acts on the complex structure of the torus. The other, sometimes called  $SL(2)_T$ , acts on the combination of the area  $\rho$  of the torus and a scalar  $b = B_{56}$  that arises in compactification of the antisymmetric tensor field  $B$ .

In addition to  $SL(2)_U$  and  $SL(2)_T$ , the heterotic string in four dimensions is conjectured to have a symmetry  $SL(2)_S$  that acts on the combination of the four-dimensional string coupling constant

$$\lambda_4 = \lambda \rho^{-1/2} \quad (4.20)$$

and a scalar  $a$  that is dual to the space-time components  $B_{mn}$  ( $m, n = 1 \dots 4$ ). We know that the heterotic string has  $SL(2)_U$  and  $SL(2)_T$  symmetry; we would like to know if it also has  $SL(2)_S$  symmetry. If so, the strong coupling behavior in four dimensions is determined.

Likewise, the six-dimensional Type IIA theory, compactified on  $\mathbf{R}^4 \times \mathbf{T}^2$ , has  $SL(2)_{U'} \times SL(2)_{T'}$  symmetry, and one would like to know if it also has  $SL(2)_{S'}$  symmetry. Here  $SL(2)_{U'}$  acts on the complex structure of the torus,  $SL(2)_{T'}$  acts on the area  $\rho'$  and scalar  $b'$  derived from  $B'_{56}$ , and  $SL(2)_{S'}$  would conjecturally act on the string coupling constant  $\lambda'_4$  and the scalar  $a'$  that is dual to the  $\mathbf{R}^4$  components of  $B'$ .

If string-string duality is correct, then the metrics in the equivalent heterotic and Type IIA descriptions differ only by a Weyl transformation, which does not change the complex structure of the torus; hence  $SL(2)_U$  can be identified with  $SL(2)_{U'}$ . More interesting is what happens to  $S$  and  $T$ . Because the duality between the heterotic string and Type IIA involves  $dB \sim *dB'$ , it turns  $a$  into  $b'$  and  $a'$  into  $b$ . Therefore, it must turn  $SL(2)_S$  into  $SL(2)_{T'}$  and  $SL(2)_{S'}$  into  $SL(2)_T$ . Hence the known  $SL(2)_T$  invariance of the heterotic

and Type IIA theories implies, if string-string duality is true, that these theories must also have  $SL(2)_S$  invariance!

It is amusing to check other manifestations of the fact that string-string duality exchanges  $SL(2)_S$  and  $SL(2)_T$ . For example, the four-dimensional string coupling  $\lambda_4 = \lambda\rho^{-1/2} = \lambda/r$  ( $r$  is a radius of the torus) turns under string-string duality into  $1/r' = (\rho')^{-1/2}$ . Likewise  $\rho = r^2$  is transformed into  $\lambda^2(r')^2 = (r'/\lambda')^2 = 1/(\lambda'_4)^2$ . So string-string duality exchanges  $\lambda_4$  with  $\rho^{-1/2}$ , as it must in order to exchange  $SL(2)_S$  and  $SL(2)_T$ .

### *Some Other Models With S-Duality*

From string-string duality we can not only rederive the familiar  $S$ -duality, but attempt to deduce  $S$ -duality for new models. For instance, one could consider in the above a particular two-torus  $\mathbf{T}^2$  that happens to be invariant under some  $SL(2)_U$  transformations, and take the orbifold with respect to that symmetry group of the six-dimensional heterotic string. This orbifold can be regarded as a different compactification of the six-dimensional model, so string-string duality – if true – can be applied to it, relating the six-dimensional heterotic string on this orbifold (and an additional four-torus) to a Type IIA string on the same orbifold (and an additional K3).

Orbifolding by a subgroup of  $SL(2)_U$  does not disturb  $SL(2)_T$ , so the basic structure used above still holds; if six-dimensional string-string duality is valid, then  $SL(2)_S$  of the heterotic string on this particular orbifold follows from  $SL(2)_T$  of Type IIA on the same orbifold, and vice-versa. This example is of some interest as – unlike previously known examples of  $S$ -duality – it involves vacua in which supersymmetry is completely broken. The  $S$ -duality of this and possible related examples might have implications in the low energy field theory limit, perhaps related to phenomena such as those recently uncovered by Seiberg [14].

#### *4.6. Enhanced Gauge Groups*

Perhaps the most striking phenomenon in toroidal compactification of the heterotic string is that at certain points in moduli space, an enhanced non-abelian gauge symmetry

appears. The enhanced symmetry group is always simply-laced and so a product of  $A$ ,  $D$ , and  $E$  groups; in toroidal compactification to six dimensions, one can get any product of  $A$ ,  $D$ , and  $E$  groups of total rank  $\leq 20$ .

How can one reproduce this with Type IIA on a K3 surface? <sup>18</sup> It is fairly obvious that one cannot get an enhanced gauge symmetry unless the K3 becomes singular; only then might the field theory analysis showing that the RR charges have mass of order  $1/\lambda$  break down.

The only singularities a K3 surface gets are orbifold singularities. (It is possible for the distance scale of the K3 to go to infinity, isotropically or not, but that just makes field theory better.) The orbifold singularities of a K3 surface have an  $A - D - E$  classification. Any combination of singularities corresponding to a product of groups with total rank  $\leq 20$  (actually at the classical level the bound is  $\leq 19$ ) can arise.

Whenever the heterotic string on a four-torus gets an enhanced gauge group  $G$ , the corresponding K3 gets an orbifold singularity of type  $G$ . This assertion must be a key to the still rather surprising and mysterious occurrence of extended gauge groups for Type IIA on K3, so I will attempt to explain it.

The moduli space

$$\mathcal{M} = SO(20, 4; \mathbf{Z}) \backslash SO(20, 4; \mathbf{R}) / (SO(20) \times SO(4)) \quad (4.21)$$

of toroidal compactifications of the heterotic string to six dimensions – or K3 compactifications of Type II – can be thought of as follows. Begin with a 24 dimensional real vector space  $W$  with a metric of signature  $(4, 20)$ , and containing a self-dual even integral lattice  $L$  (necessarily of the same signature). Let  $V$  be a four-dimensional subspace of  $W$  on which the metric of  $W$  is positive definite. Then  $\mathcal{M}$  is the space of all such  $V$ 's, up to automorphisms of  $L$ . Each  $V$  has a twenty-dimensional orthocomplement  $V^\perp$  on which the metric is negative definite.

In the heterotic string description,  $V$  is the space of charges carried by right-moving string modes, and  $V^\perp$  is the space of charges carried by left-moving string modes. Gener-

---

<sup>18</sup> This question was very briefly raised in section 4.3 of [34], and has also been raised by other physicists.

ically, neither  $V$  nor  $V^\perp$  contains any non-zero points in  $L$ . When  $V^\perp$  contains such a point  $P$ , we get a purely left-moving (antiholomorphic) vertex operator  $\mathcal{O}_P$  of dimension  $d_P = -(P, P)/2$ . (Of course,  $(P, P) < 0$  as the metric of  $W$  is negative definite on  $V^\perp$ .)  $d_P$  is always an integer as the lattice  $L$  is even. The gauge symmetry is extended precisely when  $V^\perp$  contains some  $P$  of  $d_P = 1$ ; the corresponding  $\mathcal{O}_P$  generate the extended gauge symmetry.

In the K3 description,  $W$  is the real cohomology of K3 (including  $H^0$ ,  $H^2$ , and  $H^4$  together [31]). The lattice  $L$  is the lattice of integral points.  $V$  is the part of the cohomology generated by self-dual harmonic forms. The interpretation is clearest if we restrict to K3's of large volume, where we can use classical geometry. Then  $H^0$  and  $H^4$  split off, and we can take for  $W$  the 22 dimensional space  $H^2$ , and for  $V$  the three-dimensional space of self-dual harmonic two-forms.

Consider a K3 that is developing an orbifold singularity of type  $G$ , with  $r$  being the rank of  $G$ . In the process, a configuration of  $r$  two-spheres  $S_i$  (with an intersection matrix given by the Dynkin diagram of  $G$ ) collapses to a point. These two-spheres are holomorphic (in one of the complex structures on the  $K3$ ), and the corresponding cohomology classes  $[S_i]$  have length squared  $-2$ . As they collapse, the  $[S_i]$  become anti-self-dual and thus – in the limit in which the orbifold singularity develops – they lie in  $V^\perp$ . (In fact, as  $S_i$  is holomorphic, the condition for  $[S_i]$  to be anti-self-dual is just that it is orthogonal to the Kahler class and so has zero area; thus the  $[S_i]$  lie in  $V^\perp$  when and only when the orbifold singularity appears and the  $S_i$  shrink to zero.) Conversely, the Riemann-Roch theorem can be used to prove that any point in  $V^\perp$  of length squared  $-2$  is associated with a collapsed holomorphic two-sphere.

In sum, precisely when an orbifold singularity of type  $G$  appears, there is in  $V^\perp$  an integral lattice of rank  $r$ , generated by points of length squared  $-2$ , namely the  $S_i$ ; the lattice is the weight lattice of  $G$  because of the structure of the intersection matrix of the  $S_i$ . This is the same condition on  $V^\perp$  as the one that leads to extended symmetry group  $G$  for the heterotic string. In the K3 description, one  $U(1)$  factor in the gauge group is associated with each collapsed two-sphere. These  $U(1)$ 's should make up the maximal torus of the extended gauge group.

Despite the happy occurrence of a singularity – and so possible breakdown of field theory – precisely when an extended gauge group should appear, the occurrence of extended gauge symmetry in Type IIA is still rather surprising. It must apparently mean that taking the string coupling to zero (which eliminates the RR charges) does not commute with developing an orbifold singularity (which conjecturally brings them to zero mass), and that conventional orbifold computations in string theory correspond to taking the string coupling to zero first, the opposite of what one might have guessed.

#### 4.7. Dynamics In Seven Dimensions

The reader might be struck by a lack of unity between the two parts of this paper. In sections two and three, we related Type II superstrings to eleven-dimensional supergravity. In the present section, we have presented evidence for the conjectured relation of Type II superstrings to heterotic superstrings. If both are valid, should not eleven-dimensional supergravity somehow enter in understanding heterotic string dynamics?

I will now propose a situation in which this seems to be true: the strong coupling limit of the heterotic string in seven dimensions. I will first propose an answer, and then try to deduce it from six-dimensional string-string duality.

The proposed answer is that the strong coupling limit of the heterotic string on  $\mathbf{R}^7 \times \mathbf{T}^3$  gives a theory whose low energy behavior is governed by eleven-dimensional supergravity on  $\mathbf{R}^7 \times K3$ ! The first point in favor of this is that the moduli spaces coincide. The moduli space of vacua of the heterotic string on  $\mathbf{R}^7 \times \mathbf{T}^3$  is

$$\mathcal{M} = \mathcal{M}_1 \times \mathbf{R}^+ \tag{4.22}$$

with

$$\mathcal{M}_1 = SO(19, 3; \mathbf{Z}) \backslash SO(19, 3; \mathbf{R}) / SO(19) \times SO(3). \tag{4.23}$$

Here  $\mathcal{M}_1$  is the usual Narain moduli space, and  $\mathbf{R}^+$  parametrizes the possible values of the string coupling constant. For eleven-dimensional supergravity compactified on  $\mathbf{R}^7 \times K3$ , the moduli space of vacua is simply the moduli space of Einstein metrics on  $K3$ . This does *not* coincide with the moduli space of  $(4, 4)$  conformal field theories on  $K3$ , because there is no second rank antisymmetric tensor field in eleven-dimensional supergravity.

Rather the moduli space of Einstein metrics of volume 1 on K3 is isomorphic to  $\mathcal{M}_1 = SO(19, 3; \mathbf{Z}) \backslash SO(19, 3; \mathbf{R}) / SO(19) \times SO(3)$ .<sup>19</sup> Allowing the volume to vary gives an extra factor of  $\mathbf{R}^+$ , so that the moduli space of Einstein metrics on K3 coincides with the moduli space  $\mathcal{M}$  of string vacua.

As usual, the next step is to see how the low energy effective theories match up. Relating these two theories only makes sense if large volume of eleven-dimensional supergravity (where perturbation theory is good) corresponds to strong coupling of the heterotic string. We recall that the bosonic fields of eleven-dimensional supergravity are a metric  $G$  and three-form  $A_3$  with action

$$I = \frac{1}{2} \int d^{11}x \sqrt{G} (R + |dA_3|^2) + \int A_3 \wedge dA_3 \wedge dA_3. \quad (4.24)$$

To reduce on  $\mathbf{R}^7 \times \text{K3}$ , we take the eleven-dimensional line-element to be  $ds^2 = \tilde{g}_{mn}dx^m dx^n + e^{2\gamma} h_{\alpha\beta} dy^\alpha dy^\beta$ , with  $m, n = 1 \dots 7$ ,  $\alpha, \beta = 1 \dots 4$ ; here  $\tilde{g}$  is a metric on  $\mathbf{R}^7$ ,  $h$  a fixed metric on K3 of volume 1, and  $e^\gamma$  the radius of the K3. The reduction of  $A_3$  on  $\mathbf{R}^7 \times \text{K3}$  gives on  $\mathbf{R}^7$  a three-form  $a_3$ , and 22 one-forms that we will generically call  $A$ . The eleven-dimensional Lagrangian becomes very schematically (only keeping track of the scaling with  $e^\gamma$ )

$$\int d^7x \sqrt{\tilde{g}} (e^{4\gamma} (\tilde{R} + |d\gamma|^2 + |da_3|^2) + |dA|^2). \quad (4.25)$$

To match this to the heterotic string in seven dimensions, we write  $\tilde{g} = e^{-4\gamma} g$ , with  $g$  the heterotic string metric in seven dimensions. We also make a duality transformation  $e^{6\gamma} da_3 = *dB$ , with  $B$  the two-form of the heterotic string. Then (4.25) turns into

$$\int d^7x \sqrt{g} e^{-6\gamma} (R + |d\gamma|^2 + |dB|^2 + |dA|^2). \quad (4.26)$$

The important point is that the Lagrangian scales with an overall factor of  $e^{-6\gamma}$ , similar to the overall factor of  $\lambda^{-2} = e^{-2\phi}$  in the low energy effective action of the heterotic string. Thus, to match eleven-dimensional supergravity on  $\mathbf{R}^7 \times \text{K3}$  with the heterotic string in

<sup>19</sup> This space parametrizes three-dimensional subspaces of positive metric in  $H^2(\text{K3}, \mathbf{R})$ . The subspace corresponding to a given Einstein metric on K3 consists of the part of the cohomology that is self-dual in that metric.

seven dimensions, one takes the radius of the K3 to be

$$e^\gamma = e^{\phi/3} = \lambda^{1/3}. \quad (4.27)$$

In particular, as we hoped, for  $\lambda \rightarrow \infty$ , the radius of the K3 goes to infinity, and the eleven-dimensional supergravity theory becomes weakly coupled at the length scale of the light degrees of freedom.

Now, let us try to show that this picture is a consequence of string-string duality in six dimensions. We start with the heterotic string on  $\mathbf{R}^6 \times \mathbf{S}^1 \times \mathbf{T}^3$ , where  $\mathbf{S}^1$  is a circle of radius  $r_1$ , and  $\mathbf{T}^3$  is a three-torus that will be held fixed throughout the discussion.<sup>20</sup> If  $\lambda_7$  and  $\lambda_6$  denote the heterotic string coupling constant in seven and six dimensions, respectively, then

$$\frac{1}{\lambda_6^2} = \frac{r_1}{\lambda_7^2}. \quad (4.28)$$

We want to take  $r_1$  to infinity, keeping  $\lambda_7$  fixed. That will give a heterotic string in seven dimensions. Then, after taking  $r_1$  to infinity, we consider the behavior for large  $\lambda_7$ , to get a strongly coupled heterotic string in seven dimensions.

The strategy of the analysis is of course to first dualize the theory, to a ten-dimensional Type II theory, and then see what happens to the dual theory when first  $r_1$  and then  $\lambda$  are taken large. Six-dimensional string-string duality says that for fixed  $r_1$  and  $\lambda$ , the heterotic string on  $\mathbf{R}^6 \times \mathbf{S}^1 \times \mathbf{T}^3$  is equivalent to a Type IIA superstring on  $\mathbf{R}^6 \times \mathbf{K3}$ , with the following change of variables. The six-dimensional string coupling constant  $\lambda'_6$  of the Type IIA description is

$$\lambda'_6 = \frac{1}{\lambda_6} = \frac{r_1^{1/2}}{\lambda_7}. \quad (4.29)$$

The metrics  $g$  and  $g'$  of the heterotic and Type IIA descriptions are related by

$$g = e^{2\phi} g' = \lambda_6^2 g' = \frac{\lambda_7^2}{r_1} g'. \quad (4.30)$$

In addition, the parameters of the K3 depend on  $r_1$  (and the parameters of the  $\mathbf{T}^3$ , which will be held fixed) in a way that we will now analyze.

<sup>20</sup> Generally, there are also Wilson lines on  $\mathbf{T}^3$  breaking the gauge group to a product of  $U(1)$ 's; these will be included with the parameters of the  $\mathbf{T}^3$  that are kept fixed in the discussion.

There is no unique answer, since we could always apply an  $SO(20, 4; \mathbf{Z})$  transformation to the K3. However, there is a particularly simple answer. The heterotic string compactified on  $\mathbf{S}^1 \times \mathbf{T}^3$  has 24 abelian gauge fields. As the radius  $r_1$  of the  $\mathbf{S}^1$  goes to infinity, the elementary string states carrying the 24 charges behave as follows. There is one type of charge (the momentum around the  $\mathbf{S}^1$ ) such that the lightest states carrying only that charge go to zero mass, with

$$M \sim \frac{1}{r_1}. \quad (4.31)$$

There is a second charge, the winding number around  $\mathbf{S}^1$ , such that particles carrying that charge have masses that blow up as  $r_1$ . Particles carrying only the other 22 charges have fixed masses in the limit.

Any two ways to reproduce this situation with a K3 will be equivalent up to a  $T$ -duality transformation. There is a particularly easy way to do this – take a fixed K3 and scale up the volume  $V$ , leaving fixed the “shape.” This reproduces the above spectrum with a relation between  $V$  and  $r_1$  that we will now determine.

We start with the Type IIA superstring theory in ten dimensions. The bosonic fields include the metric  $g'_{10}$ , dilaton  $\phi'_{10}$ , gauge field  $A$ , and three-form  $A_3$ . The action is schematically

$$\int d^{10}x \sqrt{g'_{10}} \left( e^{-2\phi'_{10}} R'_{10} + |dA|^2 + |dA_3|^2 + \dots \right). \quad (4.32)$$

Upon compactification on  $\mathbf{R}^6 \times \text{K3}$ , massless modes coming from  $A$  and  $A_3$  are as follows.  $A$  gives rise to a six-dimensional vector, which we will call  $a$ .  $A_3$  gives rise to 22 vectors – we will call them  $C_I$  – and a six-dimensional three-form, which we will call  $a_3$ . If  $V$  is the volume of the K3, the effective action in six dimensions scales schematically as

$$\int d^6x \sqrt{g'} \left( \frac{1}{(\lambda'_6)^2} R' + V|da|^2 + V|da_3|^2 + |dC_I|^2 \right). \quad (4.33)$$

Visible in (4.33) are 23 vectors, namely  $a$  and the  $C_I$ . However, precisely in six dimensions a three-form is dual to a vector, by  $V da_3 = *db$ . So we can replace (4.33) with

$$\int d^6x \sqrt{g'} \left( \frac{1}{(\lambda'_6)^2} R' + V|da|^2 + \frac{1}{V}|db|^2 + |dC_I|^2 \right), \quad (4.34)$$

with 24 vectors. As the canonical kinetic energy of a vector is

$$\int d^6x \frac{1}{4e_{\text{eff}}^2} |dA|^2, \quad (4.35)$$

with  $e_{\text{eff}}$  the effective charge, we see that we have one vector with effective charge of order  $V^{-1/2}$ , one with effective charge of order  $V^{1/2}$ , and 22 with effective charges of order one.

According to our discussion in section two, the mass of a particle carrying an RR charge is of order  $e_{\text{eff}}/\lambda'_6$ . So for fixed  $\lambda'_6$  and  $V \rightarrow \infty$ , one type of particle goes to zero mass, one to infinite mass, and 22 remain fixed – just like the behavior of the heterotic string as  $r_1 \rightarrow \infty$ . The highest charge-bearing particle has a mass of order

$$M' = \frac{1}{V^{1/2}\lambda'_6}. \quad (4.36)$$

To compare this to the mass (4.31) of the lightest particle in the heterotic string description, we must remember the Weyl transformation (4.30) between the two descriptions. Because of this Weyl transformation, the relation between the two masses should be  $M = \lambda_6^{-1}M' = \lambda'_6 M'$ . So  $\lambda'_6$  scales out, and the relation between the two descriptions involves the transformation

$$V = r_1^2. \quad (4.37)$$

The reason that the string coupling constant scales out is that it does not enter the map between the moduli space of heterotic string vacua on a four-torus and (4,4) conformal field theories on K3; the relation (4.37) could have been deduced by studying the description of quantum K3 moduli space in [31] instead of using low energy supergravity as we have done.

Since we know from (4.37) and (4.29) how the parameters  $V$  and  $\lambda'_6$  of the Type IIA description are related to the heterotic string parameters, we can identify the ten-dimensional Type IIA string coupling constant  $\lambda'_{10}$ , given by

$$\frac{V}{(\lambda'_{10})^2} = \frac{1}{(\lambda'_6)^2}. \quad (4.38)$$

We get

$$\lambda'_{10} = \frac{r_1^{3/2}}{\lambda_7}. \quad (4.39)$$

Thus, for  $r_1 \rightarrow \infty$ , the Type IIA theory is becoming strongly coupled. At the same time, according to (4.37) one has  $V \rightarrow \infty$ , so the Type IIA theory is becoming decompactified.

In section two, we proposed a candidate for the strong coupling behavior of Type IIA on  $\mathbf{R}^{10}$ : it is given by eleven-dimensional supergravity on  $\mathbf{R}^{10} \times \mathbf{S}^1$ . To be more precise, the relation acted as follows on the massless modes. If the line element of the eleven-dimensional theory is  $ds^2 = G_{ij}^{10} dx^i dx^j + r_{11}^2 (dx^{11})^2$ ,  $i, j = 1 \dots 10$ , with  $G^{10}$  a metric on  $\mathbf{R}^{10}$  and  $r_{11}$  the radius of the circle, then  $r_{11}$  is related to the ten-dimensional Type IIA string coupling constant by

$$r_{11} = (\lambda'_{10})^{2/3} = \frac{r_1}{\lambda_7^{2/3}} \quad (4.40)$$

and the Type IIA metric  $g'$  is related to  $G^{10}$  by

$$g' = (\lambda'_{10})^{2/3} G^{10}. \quad (4.41)$$

As this result holds for any fixed metric  $g'$  on  $\mathbf{R}^{10}$ , it must, physically, hold on any ten-manifold  $M$  as long as the dimensions of  $M$  are scaled up fast enough compared to the growth of the ten-dimensional string coupling constant. I will assume that with  $\lambda'_{10}$  and  $V$  going to infinity as determined above, one is in the regime in which one can use the formulas (4.40), (4.41) that govern the strong coupling behavior on  $\mathbf{R}^{10}$ .

If this is so, then from (4.41) the volume  $V_{11}$  of the K3 using the metric of the eleven-dimensional supergravity is related to the volume  $V$  using the string metric of the Type IIA description by

$$V_{11} = (\lambda'_{10})^{-4/3} V = \lambda_7^{4/3} r_1^{-2} V = \lambda_7^{4/3}. \quad (4.42)$$

Now we have the information we need to solve our problem. The heterotic string on  $\mathbf{R}^6 \times \mathbf{S}^1 \times \mathbf{T}^3$ , with radius  $r_1$  of the  $\mathbf{S}^1$  and string coupling constant  $\lambda_7$ , is related to eleven-dimensional supergravity on  $\mathbf{R}^6 \times \mathbf{S}^1 \times \text{K3}$ , where the radius of the  $\mathbf{S}^1$  is given in (4.40) and the volume of the K3 in (4.42). We are supposed to take the limit  $r_1 \rightarrow \infty$  and then consider the behavior for large  $\lambda_7$ . The key point is that  $V_{11}$  is independent of  $r_1$ . This enables us to take the limit as  $r_1 \rightarrow \infty$ ; all that happens is that  $r_{11} \rightarrow \infty$ , so the  $\mathbf{R}^6 \times \mathbf{S}^1 \times \text{K3}$  on which the supergravity theory is formulated becomes  $\mathbf{R}^7 \times \text{K3}$ . (Thus we see Lorentz invariance between the “eleventh” dimension which came from strong coupling and six of the “original” dimensions.) The dependence on the heterotic string coupling  $\lambda_7$  is now easy to understand: it is simply that the volume of the K3 is  $V_{11} \sim \lambda_7^{4/3}$ . That

is of course the behavior of the volume expected from (4.27). So the relation that we have proposed between the heterotic string in seven dimensions and eleven-dimensional supergravity on  $\mathbf{R}^7 \times K3$  fits very nicely with the implications of string-string duality in six dimensions.

## 5. On Heterotic String Dynamics Above Seven Dimensions

By now we have learned that the strong coupling dynamics of Type II superstrings is, apparently, tractable in any dimension and that the same appears to be true of the heterotic string in dimension  $\leq 7$ . Can we also understand the dynamics of the heterotic string above seven dimensions?

It might be possible to extend the use of six-dimensional string-string duality above seven dimensions (just as we extended it above six dimensions at the end of the last section). This will require more careful analysis of the  $K3$ 's and probably more subtle degenerations than we have needed so far.

But is there some dual description of the heterotic string above seven dimensions that would give the dynamics more directly? For instance, can we find a dual of the heterotic string directly in ten dimensions?

Once this question is asked, an obvious speculation presents itself, at least in the case of  $SO(32)$ . (For the  $E_8 \times E_8$  theory in ten dimensions, I have no proposal to make.) There is another ten-dimensional string theory with  $SO(32)$  gauge group, namely the Type I superstring. Might they in fact be equivalent?<sup>21</sup>

The low energy effective theories certainly match up; this follows just from the low energy supersymmetry. Moreover, they match up in such a way that strong coupling of one theory would turn into weak coupling of the other. This is an essential point in any possible relation between them, since weak coupling of one is certainly not equivalent to weak coupling of the other. In terms of the metric  $g$ , dilaton  $\phi$ , two-form  $B$ , and gauge field

<sup>21</sup> The  $SO(32)$  heterotic string has particles that transform as spinors of  $SO(32)$ ; these are absent in the elementary string spectrum of Type I and would have to arise as some sort of solitons if these two theories are equivalent.

strength  $F$ , the heterotic string effective action in ten dimensions scales with the dilaton like

$$\int d^{10}x \sqrt{g} e^{-2\phi} (R + |\nabla\phi|^2 + F^2 + |dB|^2). \quad (5.1)$$

If we transform  $g = e^\phi g'$  and  $\phi = -\phi'$ , this scales like

$$\int d^{10}x \sqrt{g'} (e^{-2\phi'} (R' + |\nabla\phi'|^2) + e^{-\phi'} F^2 + |dB|^2). \quad (5.2)$$

This is the correct scaling behavior for the effective action of the Type I superstring. The gauge kinetic energy scales as  $e^{-\phi'}$  instead of  $e^{-2\phi'}$  because it comes from the disc instead of the sphere. The  $B$  kinetic energy scales trivially with  $\phi'$  in Type I because  $B$  is an RR field. The fact that  $\phi = -\phi'$  means that strong coupling of one theory is weak coupling of the other, as promised.

Though a necessary condition, this is scarcely strong evidence for a new string-string duality between the heterotic string and Type I. However, given that the heterotic and Type II superstrings and eleven-dimensional supergravity all apparently link up, one would be reluctant to overlook a possibility for Type I to also enter the story.

Let us try to use this hypothetical new duality to determine the dynamics of the heterotic string below ten dimensions. (Below ten dimensions, the  $SO(32)$  and  $E_8 \times E_8$  heterotic strings are equivalent [9], so the following discussion applies to both.) We formulate the heterotic string, with ten-dimensional string coupling constant  $\lambda$ , on  $\mathbf{R}^d \times \mathbf{T}^{10-d}$  with  $\mathbf{T}^{10-d}$  a  $(10-d)$ -torus of radius  $r$ . This would be hypothetically equivalent to a toroidally compactified Type I theory with coupling constant  $\lambda' = 1/\lambda$  and (in view of the Weyl transformation used to relate the low energy actions) compactification scale  $r' = r/\lambda^{1/2}$ . Thus, as  $\lambda \rightarrow \infty$  for fixed  $r$ ,  $\lambda'$  goes to zero, but  $r'$  also goes to zero, making the physical interpretation obscure. It is more helpful to make a  $T$ -duality transformation of the Type I theory to one with radius  $r'' = 1/r'$ . The  $T$ -duality transformation has a very unusual effect for Type I superstrings [11], mapping them to a system that is actually somewhat similar to a Type II orbifold; the relation of this unusual orbifold to the system considered in section four merits further study. The  $T$ -duality transformation also changes the ten-dimensional string coupling constant to a new one  $\lambda''$  which obeys

$$\frac{(r')^{10-d}}{(\lambda')^2} = \frac{(r'')^{10-d}}{(\lambda'')^2} \quad (5.3)$$

so that the  $d$ -dimensional effective Newton constant is invariant. Thus

$$\lambda'' = \lambda' \left( \frac{r''}{r'} \right)^{(10-d)/2} = \frac{\lambda^{(8-d)/2}}{r^{10-d}}. \quad (5.4)$$

So for  $d = 9$ , the strong coupling problem would be completely solved: as  $\lambda \rightarrow \infty$  with fixed  $r$ ,  $\lambda'' \rightarrow 0$  (and  $r'' \rightarrow \infty$ , which gives further simplification). For  $d = 8$ , we have a story similar to what we have already found in  $d = 5$  and  $7$  (and for Type IIA in  $d = 10$ ): though  $\lambda''$  is of order 1, the fact that  $r'' \rightarrow \infty$  means that the coupling is weak at the compactification scale, so that one should have a weakly coupled description of the light degrees of freedom. But below  $d = 8$ , the transformation maps one strong coupling limit to another.

Of course, once we get down to seven dimensions, we have a conjecture about the heterotic string dynamics from the relation to Type II. Perhaps it is just as well that the speculative relation of the heterotic string to Type I does not give a simple answer below eight dimensions. If there were a dimension in which both approaches could be applied, then by comparing them we would get a relation between (say) a weakly coupled Type II string and a weakly coupled Type I string. Such a relation would very likely be false, so the fact that the speculative string-string duality in ten dimensions does not easily determine the strong coupling behavior below  $d = 8$  could be taken as a further (weak) hint in its favor.

I would like to thank A. Borel, D. Morrison, R. Plesser and N. Seiberg for discussions.

## References

- [1] A. Sen, “Strong-Weak Coupling Duality In Four-Dimensional String Theory,” *Int. J. Mod. Phys. **A9*** (1994) 3707, hep-th/9402002.
- [2] J. H. Schwarz, “Evidence For Non-Perturbative String Symmetries,” hep-th/9411178.
- [3] C. M. Hull and P. K. Townsend, “Unity Of Superstring Dualities,” QMW-94-30, R/94/33.
- [4] C. Vafa, unpublished.
- [5] W. Nahm, “Supersymmetries And Their Representations,” *Nucl. Phys. **B135*** (1978) 149.
- [6] E. Cremmer, B. Julia, and J. Scherk, “Supergravity Theory In 11 Dimensions,” *Phys. Lett. **76B*** (1978) 409.
- [7] M. Duff, “Strong/Weak Coupling Duality From The Dual String” hep-th/9501030.
- [8] K. Narain, “New Heterotic String Theories In Uncompactified Dimensions  $< 10$ ,” *Phys. Lett. **169B*** (1986) 41; K. Narain, M. Samadi, and E. Witten, “A Note On The Toroidal Compactification Of Heterotic String Theory,” *Nucl. Phys. **B279*** (1987) 369.
- [9] P. Ginsparg, “On Toroidal Compactification Of Heterotic Superstrings,” *Phys. Rev. **D35*** (1987) 648.
- [10] M. Dine, P. Huet, and N. Seiberg, “Large And Small Radius In String Theory,” *Nucl. Phys. **B322*** (1989) 301.
- [11] J. Dai, R. G. Leigh, and J. Polchinski, “New Connections Between String Theories,” *Mod. Phys. Lett. **A4*** (1989) 2073.
- [12] P. Townsend, “The Eleven-Dimensional Supermembrane Revisited,” hep-th-9501068.
- [13] M. F. Duff, “Duality Rotations In String Theory,” *Nucl. Phys. **B335*** (1990) 610; M. F. Duff and J. X. Lu, “Duality Rotations in Membrane Theory,” *Nucl. Phys. **B347*** (1990) 394.
- [14] N. Seiberg, “Electric-Magnetic Duality In Supersymmetric Non-Abelian Gauge Theories,” hep-th/9411149.
- [15] I. Bars, “First Massive Level And Anomalies In The Supermembrane,” *Nucl. Phys. **B308*** (1988) 462.
- [16] J. L. Carr, S. J. Gates, Jr., and R. N. Oerter, “ $D = 10$ ,  $N = 2a$  Supergravity in Superspace,” *Phys. Lett. **189B*** (1987) 68.
- [17] E. Witten and D. Olive, “Supersymmetry Algebras That Include Central Charges,” *Phys. Lett. **B78*** (1978) 97.
- [18] G. W. Gibbons and C. M. Hull, “A Bogomol’ny Bound For General Relativity And Solitons in  $N = 2$  Supergravity,” *Phys. Lett. **109B*** (1982) 190.
- [19] R. Kallosh, A. Linde, T. Ortin, A. Peet, and A. Van Proeyen, “Supersymmetry As A Cosmic Censor,” *Phys. Rev. **D46*** (1992) 5278.

- [20] S. Shenker, “The Strength Of Non-Perturbative Effects In String Theory,” in the Proceedings of the Cargese Workshop On Random Surfaces, Quantum Gravity, And Strings (1990).
- [21] N. Seiberg and E. Witten, “Electric-Magnetic Duality, Monopole Condensation, And Confinement In  $N = 2$  Supersymmetric Yang-Mills Theory,” Nucl. Phys. **B426** (1994) 19.
- [22] M. Huq and M. A. Namazie, “Kaluza-Klein Supergravity In Ten Dimensions,” Class. Quantum Grav. **2** (1985) 293.
- [23] E. Cremmer and B. Julia, “The  $SO(8)$  Supergravity,” Nucl. Phys. **B159** (1979) 141; B. Julia, “Group Disintegrations,” in *Superspace And Supergravity*, ed. M. Rocek and S. Hawking (Cambridge University Press, 1981), p. 331.
- [24] A. Salam and E. Sezgin, eds., *Supergravity In Diverse Dimensions* (North-Holland/World Scientific, 1989).
- [25] A. Sen, “Strong-Weak Coupling Duality In Three-Dimensional String Theory,” hepth/9408083.
- [26] A. Dabholkar and J. A. Harvey, “Nonrenormalization Of The Superstring Tension,” Phys. Rev. Lett. **63** (1989) 719; A. Dabholkar, G. Gibbons, J. A. Harvey, and F. Ruiz Ruiz, “Superstrings And Solitons,” Nucl. Phys. **B340** (1990) 33.
- [27] A. Strominger, “Heterotic Solitons,” Nucl. Phys. **B343** (1990) 167.
- [28] C. G. Callan, Jr., J. A. Harvey, and A. Strominger, “World-Sheet Approach To Heterotic Instantons And Solitons,” Nucl. Phys. **B359** (1991) 611.
- [29] J. Gauntlett and J. A. Harvey “ $S$ -Duality And The Spectrum Of Magnetic Monopoles In Heterotic String Theory,” hepth/9407111.
- [30] N. Seiberg, “Observations On The Moduli Space of Superconformal Field Theories,” Nucl. Phys. **B303** (1988) 286
- [31] P. Aspinwall and D. Morrison, “String Theory On K3 Surfaces,” DUK-TH-94-68, IASSNS-HEP-94/23.
- [32] L. Romans, “Self-Duality For Interacting Fields: Covariant Field Equations For Six-Dimensional Chiral Supergravities,” Nucl. Phys. **B276** (1986) 71.
- [33] A. Giveon, M. Petratti, and E. Rabinovici, “Target Space Duality In String Theory,” hepth/9401139.
- [34] A. Ceresole, R. D’Auria, S. Ferrara, and A. Van Proeyen, “Duality Transformations In Supersymmetric Yang-Mills Theories Coupled To Supergravity,” hepth-9502072

# The Large N Limit of Superconformal field theories and supergravity

Juan Maldacena<sup>1</sup>

*Lyman Laboratory of Physics, Harvard University, Cambridge, MA 02138, USA*

## Abstract

We show that the large  $N$  limit of certain conformal field theories in various dimensions include in their Hilbert space a sector describing supergravity on the product of Anti-deSitter spacetimes, spheres and other compact manifolds. This is shown by taking some branes in the full M/string theory and then taking a low energy limit where the field theory on the brane decouples from the bulk. We observe that, in this limit, we can still trust the near horizon geometry for large  $N$ . The enhanced supersymmetries of the near horizon geometry correspond to the extra supersymmetry generators present in the superconformal group (as opposed to just the super-Poincare group). The 't Hooft limit of 3+1  $\mathcal{N} = 4$  super-Yang-Mills at the conformal point is shown to contain strings: they are IIB strings. We conjecture that compactifications of M/string theory on various Anti-deSitter spacetimes is dual to various conformal field theories. This leads to a new proposal for a definition of M-theory which could be extended to include five non-compact dimensions.

---

<sup>1</sup> malda@pauli.harvard.edu

## 1. General idea

In the last few years it has been extremely fruitful to derive quantum field theories by taking various limits of string or M-theory. In some cases this is done by considering the theory at geometric singularities and in others by considering a configuration containing branes and then taking a limit where the dynamics on the brane decouples from the bulk. In this paper we consider theories that are obtained by decoupling theories on branes from gravity. We focus on conformal invariant field theories but a similar analysis could be done for non-conformal field theories. The cases considered include  $N$  parallel D3 branes in IIB string theory and various others. We take the limit where the field theory on the brane decouples from the bulk. At the same time we look at the near horizon geometry and we argue that the supergravity solution can be trusted as long as  $N$  is large.  $N$  is kept fixed as we take the limit. The approach is similar to that used in [1] to study the NS fivebrane theory [2] at finite temperature. The supergravity solution typically reduces to  $p+2$  dimensional Anti-deSitter space ( $AdS_{p+2}$ ) times spheres (for D3 branes we have  $AdS_5 \times S^5$ ). The curvature of the sphere and the  $AdS$  space in Planck units is a (positive) power of  $1/N$ . Therefore the solutions can be trusted as long as  $N$  is large. Finite temperature configurations in the decoupled field theory correspond to black hole configurations in  $AdS$  spacetimes. These black holes will Hawking radiate into the  $AdS$  spacetime. We conclude that excitations of the  $AdS$  spacetime are included in the Hilbert space of the corresponding conformal field theories. A theory in  $AdS$  spacetime is not completely well defined since there is a horizon and it is also necessary to give some boundary conditions at infinity. However, local properties and local processes can be calculated in supergravity when  $N$  is large if the proper energies involved are much bigger than the energy scale set by the cosmological constant (and smaller than the Planck scale). We will conjecture that the full quantum M/string-theory on  $AdS$  space, plus suitable boundary conditions is dual to the corresponding brane theory. We are not going to specify the boundary conditions in  $AdS$ , we leave this interesting problem for the future. The  $AdS \times (\text{spheres})$  description will become useful for large  $N$ , where we can isolate some local processes from the question of boundary conditions. The supersymmetries of both theories agree, both are given by the superconformal group. The superconformal group has twice the amount of supersymmetries of the corresponding super-Poincare group[3,4]. This enhancement of supersymmetry near the horizon of extremal black holes was observed in [5,6] precisely by showing that the near throat geometry reduces to  $AdS \times (\text{spheres})$ .  $AdS$  spaces (and

branes in them) were extensively considered in the literature [7,8,9,10,11,12,13], includding the connection with the superconformal group.

In section 2 we study  $\mathcal{N} = 4$  d=4  $U(N)$  super-Yang-Mills as a first example, we discuss several issues which are present in all other cases. In section 3 we analyze the theories describing M-theory five-branes and M-theory two-branes. In section 4 we consider theories with lower supersymmetry which are related to a black string in six dimensions made with D1 and D5 branes. In section 5 we study theories with even less supersymmetry involving black strings in five dimensions and finally we mention the theories related to extremal Reissner-Nordström black holes in four spacetime dimensions (these last cases will be more speculative and contain some unresolved puzzles). Finally in section 6 we make some comments on the relation to matrix theory.

## 2. D3 branes or $\mathcal{N} = 4$ $U(N)$ super-Yang-Mills in d=3+1

We start with type IIB string theory with string coupling  $g$ , which will remain fixed. Consider  $N$  parallel D3 branes separated by some distances which we denote by  $r$ . For low energies the theory on the D3 brane decouples from the bulk. It is more convenient to take the energies fixed and take

$$\alpha' \rightarrow 0 , \quad U \equiv \frac{r}{\alpha'} = \text{fixed} . \quad (2.1)$$

The second condition is saying that we keep the mass of the stretched strings fixed. As we take the decoupling limit we bring the branes together but the the Higgs expectation values corresponding to this separation remains fixed. The resulting theory on the brane is four dimensional  $\mathcal{N} = 4$   $U(N)$  SYM. Let us consider the theory at the superconformal point, where  $r = 0$ . The conformal group is  $SO(2,4)$ . We also have an  $SO(6) \sim SU(4)$  R-symmetry that rotates the six scalar fields into each other<sup>2</sup>. The superconformal group includes twice the number of supersymmetries of the super-Poincare group: the commutator of special conformal transformations with Poincare supersymmetry generators gives the new supersymmetry generators. The precise superconformal algebra was computed in [3]. All this is valid for any  $N$ .

---

<sup>2</sup> The representation includes objects in the spinor representations, so we should be talking about  $SU(4)$ , we will not make this, or similar distinctions in what follows.

Now we consider the supergravity solution carrying D3 brane charge [14]

$$ds^2 = f^{-1/2} dx_{||}^2 + f^{1/2} (dr^2 + r^2 d\Omega_5^2) , \\ f = 1 + \frac{4\pi g N \alpha'^2}{r^4} , \quad (2.2)$$

where  $x_{||}$  denotes the four coordinates along the worldvolume of the three-brane and  $d\Omega_5^2$  is the metric on the unit five-sphere<sup>3</sup>. The self dual five-form field strength is nonzero and has a flux on the five-sphere. Now we define the new variable  $U \equiv \frac{r}{\alpha'}$  and we rewrite the metric in terms of  $U$ . Then we take the  $\alpha' \rightarrow 0$  limit. Notice that  $U$  remains fixed. In this limit we can neglect the 1 in the harmonic function (2.2). The metric becomes

$$ds^2 = \alpha' \left[ \frac{U^2}{\sqrt{4\pi g N}} dx_{||}^2 + \sqrt{4\pi g N} \frac{dU^2}{U^2} + \sqrt{4\pi g N} d\Omega_5^2 \right] . \quad (2.3)$$

This metric describes five dimensional Anti-deSitter ( $AdS_5$ ) times a five-sphere<sup>4</sup>. We see that there is an overall  $\alpha'$  factor. The metric remains constant in  $\alpha'$  units. The radius of the five-sphere is  $R_{sph}^2/\alpha' = \sqrt{4\pi g N}$ , and is the same as the “radius” of  $AdS_5$  (as defined in the appendix). In ten dimensional Planck units they are both proportional to  $N^{1/4}$ . The radius is quantized because the flux of the 5-form field strength on the 5 sphere is quantized. We can trust the supergravity solution when

$$gN \gg 1 . \quad (2.4)$$

When  $N$  is large we have approximately ten dimensional flat space in the neighborhood of any point<sup>5</sup>. Note that in the large  $N$  limit the flux of the 5 form field strength per unit Planck (or string) 5-volume becomes small.

Now consider a near extremal black D3 brane solution in the decoupling limit (2.1). We keep the energy density on the brane worldvolume theory ( $\mu$ ) fixed. We find the metric

$$ds^2 = \alpha' \left\{ \frac{U^2}{\sqrt{4\pi g N}} [-(1 - U_0^4/U^4) dt^2 + dx_i^2] + \sqrt{4\pi g N} \frac{dU^2}{U^2(1 - U_0^4/U^4)} + \sqrt{4\pi g N} d\Omega_5^2 \right\} . \\ U_0^4 = \frac{2^7}{3} \pi^4 g^2 \mu \quad (2.5)$$

<sup>3</sup> We choose conventions where  $g \rightarrow 1/g$  under S-duality.

<sup>4</sup> See the appendix for a brief description of  $AdS$  spacetimes.

<sup>5</sup> In writing (2.4) we assumed that  $g \leq 1$ , if  $g > 1$  then the condition is  $N/g \gg 1$ . In other words we need large  $N$ , not large  $g$ .

We see that  $U_0$  remains finite when we take the  $\alpha' \rightarrow 0$  limit. The situation is similar to that encountered in [1]. Naively the whole metric is becoming of zero size since we have a power of  $\alpha'$  in front of the metric, and we might incorrectly conclude that we should only consider the zero modes of all fields. However, energies that are finite from the point of view of the gauge theory, lead to proper energies (measured with respect to proper time) that remain finite in  $\alpha'$  units (or Planck units, since  $g$  is fixed). More concretely, an excitation that has energy  $\omega$  (fixed in the limit) from the point of view of the gauge theory, will have proper energy  $E_{proper} = \frac{1}{\sqrt{\alpha'}} \frac{\omega(gN4\pi)^{1/4}}{U}$ . This also means that the corresponding proper wavelengths remain fixed. In other words, the spacetime action on this background has the form  $S \sim \frac{1}{\alpha'^4} \int d^{10}x \sqrt{G}R + \dots$ , so we can cancel the factor of  $\alpha'$  in the metric and the Newton constant, leaving a theory with a finite Planck length in the limit. Therefore we should consider fields that propagate on the  $AdS$  background. Since the Hawking temperature is finite, there is a flux of energy from the black hole to the  $AdS$  spacetime. Since  $\mathcal{N} = 4$  d=4  $U(N)$  SYM is a unitary theory we conclude that, for large  $N$ , *it includes in its Hilbert space the states of type IIB supergravity on  $(AdS_5 \times S_5)_N$* , where subscript indicates the fact that the “radii” in Planck units are proportional to  $N^{1/4}$ . In particular the theory contains gravitons propagating on  $(AdS_5 \times S_5)_N$ . When we consider supergravity on  $AdS_5 \times S_5$ , we are faced with global issues like the presence of a horizon and the boundary conditions at infinity. It is interesting to note that the solution is nonsingular [15]. The gauge theory should provide us with a specific choice of boundary conditions. It would be interesting to determine them.

We have started with a quantum theory and we have seen that it includes gravity so it is natural to think that this correspondence goes beyond the supergravity approximation. We are led to the conjecture that *Type IIB string theory on  $(AdS_5 \times S^5)_N$  plus some appropriate boundary conditions (and possibly also some boundary degrees of freedom) is dual to  $N = 4$  d=3+1  $U(N)$  super-Yang-Mills*. The SYM coupling is given by the (complex) IIB string coupling, more precisely  $\frac{1}{g_{YM}^2} + i \frac{\theta}{8\pi^2} = \frac{1}{2\pi} (\frac{1}{g} + i \frac{\chi}{2\pi})$  where  $\chi$  is the value of the RR scalar.

The supersymmetry group of  $AdS_5 \times S^5$ , is known to be the same as the superconformal group in 3+1 spacetime dimensions [3], so the supersymmetries of both theories are the same. This is a new form of “duality”: a large  $N$  field theory is related to a string theory on some background, notice that the correspondence is non-perturbative in  $g$  and the  $SL(2, \mathbb{Z})$  symmetry of type IIB would follow as a consequence of the  $SL(2, \mathbb{Z})$  symmetry

of SYM<sup>6</sup>. It is also a strong-weak coupling correspondence in the following sense. When the effective coupling  $gN$  becomes large we cannot trust perturbative calculations in the Yang-Mills theory but we can trust calculations in supergravity on  $(AdS_5 \times S^5)_N$ . This is suggesting that the  $\mathcal{N} = 4$  Yang-Mills master field is the anti-deSitter supergravity solution (similar ideas were suggested in [17]). Since  $N$  measures the size of the geometry in Planck units, we see that quantum effects in  $AdS_5 \times S^5$  have the interpretation of  $1/N$  effects in the gauge theory. So Hawking radiation is a  $1/N$  effect. It would be interesting to understand more precisely what the horizon means from the gauge theory point of view. IIB supergravity on  $AdS_5 \times S^5$  was studied in [7,9].

The above conjecture becomes nontrivial for large  $N$  and gives a way to answer some large  $N$  questions in the SYM theory. For example, suppose that we break  $U(N) \rightarrow U(N-1) \times U(1)$  by Higgsing. This corresponds to putting a three brane at some point on the 5-sphere and some value of  $U$ , with world volume directions along the original four dimensions ( $x_{||}$ ). We could now ask what the low energy effective action for the light U(1) fields is. For large  $N$  (2.4) it is the action of a D3 brane in  $AdS_5 \times S^5$ . More concretely, the bosonic part of the action becomes the Born-Infeld action on the  $AdS$  background

$$S = -\frac{1}{(2\pi)^3 g} \int d^4x h^{-1} \left[ \sqrt{-\text{Det}(\eta_{\alpha\beta} + h\partial_\alpha U \partial_\beta U + U^2 h g_{ij} \partial_\alpha \theta^i \partial_\beta \theta^j + 2\pi\sqrt{h} F_{\alpha\beta})} - 1 \right]$$

$$h = \frac{4\pi g N}{U^4}, \quad (2.6)$$

with  $\alpha, \beta = 0, 1, 2, 3$ ,  $i, j = 1, \dots, 5$ ; and  $g_{ij}$  is the metric of the unit five-sphere. As any low energy action, (2.6) is valid when the energies are low compared to the mass of the massive states that we are integrating out. In this case the mass of the massive states is proportional to  $U$  (with no factors of  $N$ ). The low energy condition translates into  $\partial U/U \ll U$  and  $\partial\theta^i \ll U$ , etc.. So the nonlinear terms in the action (2.6) will be important only when  $gN$  is large. It seems that the form of this action is completely determined by superconformal invariance, by using the broken and unbroken supersymmetries, in the same sense that the Born Infeld action in flat space is given by the full Poincare supersymmetry [18]. It would be very interesting to check this explicitly. We will show this for a particular term in the action. We set  $\theta^i = \text{const}$  and  $F = 0$ , so that we only have  $U$  left. Then we will show that the action is completely determined by broken conformal invariance. This can

---

<sup>6</sup> This is similar in spirit to [16] but here  $N$  is not interpreted as momentum.

be seen as follows. Using Lorentz invariance and scaling symmetry (dimensional analysis) one can show that the action must have the form

$$S = \int d^{p+1}x U^{p+1} f(\partial_\alpha U \partial^\alpha U / U^4) , \quad (2.7)$$

where  $f$  is an arbitrary function. Now we consider infinitesimal special conformal transformations

$$\begin{aligned} \delta x^\alpha &= \epsilon^\beta x_\beta x^\alpha - \epsilon^\alpha (x^2 + \frac{\tilde{R}^4}{U^2})/2 , \\ \delta U &\equiv U'(x') - U(x) = -\epsilon^\alpha x_\alpha U , \end{aligned} \quad (2.8)$$

where  $\epsilon^\alpha$  is an infinitesimal parameter. For the moment  $\tilde{R}$  is an arbitrary constant. We will later identify it with the “radius” of  $AdS$ , it will turn out that  $\tilde{R}^4 \sim gN$ . In the limit of small  $\tilde{R}$  we recover the more familiar form of the conformal transformations ( $U$  is a weight one field). Usually conformal transformations do not involve the variable  $U$  in the transformations of  $x$ . For constant  $U$  the extra term in (2.8) is a translation in  $x$ , but we will take  $U$  to be a slowly varying function of  $x$  and we will determine  $\tilde{R}$  from other facts that we know. Demanding that (2.7) is invariant under (2.8) we find that the function  $f$  in (2.7) obeys the equation

$$f(z) + const = 2 \left( z + \frac{1}{\tilde{R}^4} \right) f'(z) \quad (2.9)$$

which is solved by  $f = b[\sqrt{1 + \tilde{R}^4 z} - a]$ . Now we can determine the constants  $a, b, \tilde{R}$  from supersymmetry. We need to use three facts. The first is that there is no force (no vacuum energy) for a constant  $U$ . This implies  $a = 1$ . The second is that the  $\partial U^2$  term ( $F^2$  term) in the  $U(1)$  action is not renormalized. The third is that the only contribution to the  $(\partial U)^4$  term (an  $F^4$  term) comes from a one loop diagram [19]. This determines all the coefficients to be those expected from (2.6) including the fact that  $\tilde{R}^4 = 4\pi gN$ . It seems very plausible that using all 32 supersymmetries we could fix the action (2.6) completely. This would be saying that (2.6) is a consequence of the symmetries and thus not a prediction<sup>7</sup>. However we can make very nontrivial predictions (though we were not able to check them). For example, if we take  $g$  to be small (but  $N$  large) we can predict that the Yang-Mills theory contains strings. More precisely, in the limit  $g \rightarrow 0$ ,  $gN = \text{fixed} \gg 1$  ('t Hooft limit) we

---

<sup>7</sup> Notice that the action (2.6) includes a term proportional to  $v^6$  similar to that calculated in [20]. Conformal symmetry explains the agreement that they would have found if they had done the calculation for 3+1 SYM as opposed to 0+1.

find free strings in the spectrum, they are IIB strings moving in  $(AdS_5 \times S^5)_{gN}$ .<sup>8</sup> The sense in which these strings are present is rather subtle since there is no energy scale in the Yang-Mills to set their tension. In fact one should translate the mass of a string state from the  $AdS$  description to the Yang-Mills description. This translation will involve the position  $U$  at which the string is sitting. This sets the scale for its mass. As an example, consider again the D-brane probe (Higgsed configuration) which we described above. From the type IIB description we expect open strings ending on the D3 brane probe. From the point of view of the gauge theory these open strings have energies  $E = \frac{U}{(4\pi g N)^{1/4}} \sqrt{N_{open}}$  where  $N_{open}$  is the integer characterizing the massive open string level. In this example we see that  $\alpha'$  disappears when we translate the energies and is replaced by  $U$ , which is the energy scale of the Higgs field that is breaking the symmetry.

Now we turn to the question of the physical interpretation of  $U$ .  $U$  has dimensions of mass. It seems natural to interpret motion in  $U$  as moving in energy scales, going to the IR for small  $U$  and to the UV for large  $U$ . For example, consider a D3 brane sitting at some position  $U$ . Due to the conformal symmetry, all physics at energy scales  $\omega$  in this theory is the same as physics at energies  $\omega' = \lambda\omega$ , with the brane sitting at  $U' = \lambda U$ .

Now let us turn to another question. We could separate a group of D3 branes from the point where they were all sitting originally. Fortunately, for the extremal case we can find a supergravity solution describing this system. All we have to do is the replacement

$$\frac{N}{U^4} \rightarrow \frac{N - M}{U^4} + \frac{M}{|\vec{U} - \vec{W}|^4}, \quad (2.10)$$

where  $\vec{W} = \vec{r}/\alpha'$  is the separation. It is a vector because we have to specify a point on  $S^5$  also. The resulting metric is

$$ds^2 = \alpha' \left[ U^2 \frac{1}{\sqrt{4\pi g} \left( N - M + \frac{MU^4}{|\vec{U} - \vec{W}|^4} \right)^{1/2}} dx_{||}^2 + \sqrt{4\pi g} \frac{1}{U^2} \left( N - M + \frac{MU^4}{|\vec{U} - \vec{W}|^4} \right)^{1/2} d\vec{U}^2 \right]. \quad (2.11)$$

---

<sup>8</sup> In fact, Polyakov [21] recently proposed that the string theory describing bosonic Yang-Mills has a new dimension corresponding to the Liouville mode  $\varphi$ , and that the metric at  $\varphi = 0$  is zero due to a “zig-zag” symmetry. In our case we see that the physical distances along the directions of the brane contract to zero as  $U \rightarrow 0$ . The details are different, since we are considering the  $\mathcal{N} = 4$  theory.

For large  $U \gg |W|$  we find basically the solution for  $(AdS_5 \times S^5)_N$  which is interpreted as saying that for large energies we do not see the fact that the conformal symmetry was broken, while for small  $U \ll |W|$  we find just  $(AdS_5 \times S^5)_{N-M}$ , which is associated to the CFT of the unbroken  $U(N-M)$  piece. Furthermore, if we consider the region  $|\vec{U} - \vec{W}| \ll |\vec{W}|$  we find  $(AdS_5 \times S^5)_M$ , which is described by the CFT of the  $U(M)$  piece.

We could in principle separate all the branes from each other. For large values of  $U$  we would still have  $(AdS_5 \times S^5)_N$ , but for small values of  $U$  we would not be able to trust the supergravity solution, but we naively get  $N$  copies of  $(AdS_5 \times S^5)_1$  which should correspond to the  $U(1)^N$ .

Now we discuss the issue of compactification. We want to consider the YM theory compactified on a torus of radii  $R_i$ ,  $x_i \sim x_i + 2\pi R_i$ , which stay fixed as we take the  $\alpha' \rightarrow 0$  limit. Compactifying the theory breaks conformal invariance and leaves only the Poincare supersymmetries. However one can still find the supergravity solutions and follow the above procedure, going near the horizon, etc. The  $AdS$  piece will contain some identifications. So we will be able to trust the supergravity solution as long as the physical length of these compact circles stays big in  $\alpha'$  units. This implies that we can trust the supergravity solution as long as we stay far from the horizon (at  $U = 0$ )

$$U \gg \frac{(gN)^{1/4}}{R_i} , \quad (2.12)$$

for all  $i$ . This is a larger bound than the naive expectation  $(1/R_i)$ . If we were considering near extremal black holes we would require that  $U_0$  in (2.5) satisfies (2.12), which is, of course, the same condition on the temperature gotten in [22].

The relation of the three-brane supergravity solution and the Yang-Mills theory has been studied in [23,24,25,26]. All the calculations have been done for near extremal D3 branes fall into the category described above. In particular the absorption cross section of the dilaton and the graviton have been shown to agree with what one would calculate in the YM theory [24,25]. It has been shown in [26] that some of these agreements are due to non-renormalization theorems for  $\mathcal{N} = 4$  YM. The black hole entropy was compared to the *perturbative* YM calculation and it agrees up to a numerical factor [23]. This is not in disagreement with the correspondence we were suggesting, It is expected that large  $gN$  effects change this numerical factor, this problem remains unsolved.

Finally notice that the group  $SO(2, 4) \times SO(6)$  suggests a twelve dimensional realization in a theory with two times [27].

### 3. Other cases with $16 \rightarrow 32$ supersymmetries, M5 and M2 brane theories

Basically all that we have said for the D3 brane carries over for the other conformal field theories describing coincident M-theory fivebranes and M-theory twobrane. We describe below the limits that should be taken in each of the two cases. Similar remarks can be made about the entropies [28], and the determination of the probe actions using superconformal invariance. Eleven dimensional supergravity on the corresponding  $AdS$  spaces was studied in [8,10,11,15].

#### 3.1. M 5 brane

The decoupling limit is obtained by taking the 11 dimensional Planck length to zero,  $l_p \rightarrow 0$ , keeping the worldvolume energies fixed and taking the separations  $U^2 \equiv r/l_p^3 =$  fixed [29]. This last condition ensures that the membranes stretching between fivebranes give rise to strings with finite tension.

The metric is<sup>9</sup>

$$ds^2 = f^{-1/3} dx_{||}^2 + f^{2/3} (dr^2 + r^2 d\Omega_4^2) , \\ f = 1 + \frac{\pi N l_p^3}{r^3} , \quad (3.1)$$

We also have a flux of the four-form field strength on the four-sphere (which is quantized). Again, in the limit we obtain

$$ds^2 = l_p^2 \left[ \frac{U^2}{(\pi N)^{1/3}} dx_{||}^2 + 4(\pi N)^{2/3} \frac{dU^2}{U^2} + (\pi N)^{2/3} d\Omega_4^2 \right] , \quad (3.2)$$

where now the “radii” of the sphere and the  $AdS_7$  space are  $R_{sph} = R_{AdS}/2 = l_p(\pi N)^{1/3}$ . Again, the “radii” are fixed in Planck units as we take  $l_p \rightarrow 0$ , and supergravity can be applied if  $N \gg 1$ .

Reasoning as above we conclude that this theory contains seven dimensional Anti-deSitter times a four-sphere, which for large  $N$  looks locally like eleven dimensional Minkowski space.

This gives us a method to calculate properties of the large  $N$  limit of the six dimensional (0,2) conformal field theory [30]. The superconformal group again coincides with the algebra of the supersymmetries preserved by  $AdS_7 \times S^4$ . The bosonic symmetries are  $SO(2,6) \times SO(5)$  [4]. We can do brane probe calculations, thermodynamic calculations [28], etc.

The conjecture is now that *the (0,2) conformal field theory is dual to M-theory on  $(AdS_7 \times S^4)_N$* , the subindex indicates the dependence of the “radius” with  $N$ .

---

<sup>9</sup> In our conventions the relation of the Planck length to the 11 dimensional Newton constant is  $G_N^{11} = 16\pi^7 l_p^9$ .

### 3.2. M2 brane

We now take the limit  $l_p \rightarrow 0$  keeping  $U^{1/2} \equiv r/l_p^{3/2}$  = fixed. This combination has to remain fixed because the scalar field describing the motion of the twobrane has scaling dimension 1/2. Alternatively we could have derived this conformal field theory by taking first the field theory limit of D2 branes in string theory as in [31,32,33], and then taking the strong coupling limit of that theory to get to the conformal point as in [34,35,36]. The fact that the theories obtained in this fashion are the same can be seen as follows. The D2 brane gauge theory can be obtained as the limit  $\alpha' \rightarrow 0$ , keeping  $g_{YM}^2 \sim g/\alpha'$  = fixed. This is the same as the limit of M-theory two branes in the limit  $l_p \rightarrow 0$  with  $R_{11}/l_p^{3/2} \sim g_{YM}$  = fixed. This is a theory where one of the Higgs fields is compact. Taking  $R^{11} \rightarrow \infty$  we see that we get the theory of coincident M2 branes, in which the SO(8) R-symmetry has an obvious origin.

The metric is

$$ds^2 = f^{-2/3} dx_{||}^2 + f^{1/3} (dr^2 + r^2 d\Omega_7^2) , \\ f = 1 + \frac{2^5 \pi^2 N l_p^6}{r^6} , \quad (3.3)$$

and there is a nonzero flux of the dual of the four-form field strength on the seven-sphere. In the decoupling limit we obtain  $AdS_4 \times S^7$ , and the supersymmetries work out correctly. The bosonic generators are given by  $SO(2, 3) \times SO(8)$ . In this case the “radii” of the sphere and  $AdS_4$  are  $R_{sph} = 2R_{AdS} = l_p(2^5 \pi^2 N)^{1/6}$ .

The entropy of the near extremal solution agrees with the expectation from dimensional analysis for a conformal theory in 2+1 dimensions [28], but the  $N$  dependence or the numerical coefficients are not understood.

Actually for the case of the two brane the conformal symmetry was used to determine the  $v^4$  term in the probe action [37], we are further saying that conformal invariance determines it to all orders in the velocity of the probe. Furthermore the duality we have proposed with M-theory on  $AdS_4 \times S^7$  determines the precise numerical coefficient.

When M-theory is involved the dimensionalities of the groups are suggestive of a thirteen dimensional realization [38].

## 4. Theories with $8 \rightarrow 16$ supersymmetries, the D1+D5 system

Now we consider IIB string theory compactified on  $M^4$  (where  $M^4 = T^4$  or  $K3$ ) to six spacetime dimensions. As a first example let us start with a D-fivebrane with four

dimensions wrapping on  $M^4$  giving a string in six dimensions. Consider a system with  $Q_5$  fivebranes and  $Q_1$  D-strings, where the D-string is parallel to the string in six dimensions arising from the fivebrane. This system is described at low energies by a 1+1 dimensional (4,4) superconformal field theory. So we take the limit

$$\alpha' \rightarrow 0 , \quad \frac{r}{\alpha'} = \text{fixed} , \quad v \equiv \frac{V_4}{(2\pi)^4 \alpha'^2} = \text{fixed} , \quad g_6 = \frac{g}{\sqrt{v}} = \text{fixed} \quad (4.1)$$

where  $V_4$  is the volume of  $M^4$ . All other moduli of  $M^4$  remain fixed. This is just a low energy limit, we keep all dimensionless moduli fixed. As a six dimensional theory, IIB on  $M^4$  contains strings. They transform under the U-duality group and they carry charges given by a vector  $q^I$ . In general we can consider a configuration where  $q^2 = \eta_{IJ} q^I q^J \neq 0$  (the metric is the U-duality group invariant), and then take the limit (4.1).

This theory has a branch which we will call the Higgs branch and one which we call the Coulomb branch. On the Higgs branch the 1+1 dimensional vector multiplets have zero expectation value and the Coulomb branch is the other one. Notice that the expectation values of the vector multiplets in the Coulomb branch remain fixed as we take the limit  $\alpha' \rightarrow 0$ .

The Higgs branch is a SCFT with (4,4) supersymmetry. This is the theory considered in [39]. The above limit includes also a piece of the Coulomb branch, since we can separate the branes by a distance such that the mass of stretched strings remains finite.

Now we consider the supergravity solution corresponding to D1+D5 branes [40]

$$ds^2 = f_1^{-1/2} f_5^{-1/2} dx_{||}^2 + f_1^{1/2} f_5^{1/2} (dr^2 + r^2 d\Omega_3^2) , \\ f_1 = \left( 1 + \frac{g\alpha' Q_1}{vr^2} \right) , \quad f_5 = \left( 1 + \frac{g\alpha' Q_5}{r^2} \right) , \quad (4.2)$$

where  $dx_{||}^2 = -dt^2 + dx^2$  and  $x$  is the coordinate along the D-string. Some of the moduli of  $M^4$  vary over the solution and attain a fixed value at the horizon which depends only on the charges and some others are constant throughout the solution. The three-form RR-field strength is also nonzero.

In the decoupling limit (4.1) we can neglect the 1's in  $f_i$  in (4.2) and the metric becomes

$$ds^2 = \alpha' \left[ \frac{U^2}{g_6 \sqrt{Q_1 Q_5}} dx_{||}^2 + g_6 \sqrt{Q_1 Q_5} \frac{dU^2}{U^2} + g_6 \sqrt{Q_1 Q_5} d\Omega_3^2 \right] . \quad (4.3)$$

The compact manifold  $M^4(Q)$  that results in the limit is determined as follows. Some of its moduli are at their fixed point value which depends only on the charges and not on

the asymptotic value of those moduli at infinity (the notation  $M^4(Q)$  indicates the charge dependence of the moduli)[41]<sup>10</sup>. The other moduli, that were constant in the black hole solution, have their original values. For example, the volume of  $M^4$  has the fixed point value  $v_{fixed} = Q_1/Q_5$ , while the six dimensional string coupling  $g_6$  has the original value. Notice that there is an overall factor of  $\alpha'$  in (4.3) which can be removed by canceling it with the factor of  $\alpha'$  in the Newton constant as explained above. We can trust the supergravity solution if  $Q_1, Q_5$  are large,  $g_6 Q_i \gg 1$ . Notice that we are talking about a six dimensional supergravity solution since the volume of  $M^4$  is a constant in Planck units (we keep the  $Q_1/Q_5$  ratio fixed). The metric (4.3) describes three dimensional  $AdS_3$  times a 3-sphere. The supersymmetries work out correctly, starting from the 8 Poincare supersymmetries we enhance then to 16 supersymmetries. The bosonic component is  $SO(2, 2) \times SO(4)$ . In conformal field theory language  $SO(2, 2)$  is just the  $SL(2, R) \times SL(2, R)$  part of the conformal group and  $SO(4) \sim SU(2)_L \times SU(2)_R$  are the R-symmetries of the CFT [43].

So the conjecture is that *the 1+1 dimensional CFT describing the Higgs branch of the D1+D5 system on  $M^4$  is dual to type IIB string theory on  $(AdS_3 \times S^3)_{Q_1 Q_5} \times M^4(Q)$* . The subscript indicates that the radius of the three sphere is  $R_{sph}^2 = \alpha' g_6 \sqrt{Q_1 Q_5}$ . The compact fourmanifold  $M^4(Q)$  is at some particular point in moduli space determined as follows. The various moduli of  $M^4$  are divided as tensors and hypers according to the (4,4) supersymmetry on the brane. Each hypermultiplet contains four moduli and each tensor contains a modulus and an anti-self-dual  $B$ -field. (There are five tensors of this type for  $T^4$  and 21 for  $K3$ ). The scalars in the tensors have fixed point values at the horizon of the black hole, and those values are the ones entering in the definition of  $M^4(Q)$  ( $Q$  indicates the dependence on the charges). The hypers will have the same expectation value everywhere. It is necessary for this conjecture to work that the 1+1 dimensional (4,4) theory is independent of the tensor moduli appearing in its original definition as a limit of the brane configurations, since  $M^4(Q)$  does not depend on those moduli. A non renormalization theorem like [44,45] would explain this. We also need that the Higgs branch decouples from the Coulomb branch as in [46,47].

Finite temperature configurations in the 1+1 conformal field theory can be considered. They correspond to near extremal black holes in  $AdS_3$ . The metric is the same as that of

---

<sup>10</sup> The fixed values of the moduli are determined by the condition that they minimize the tension of the corresponding string (carrying charges  $q^I$ ) in six dimensions [41]. This is parallel to the discussion in four dimensions [42].

the BTZ 2+1 dimensional black hole [48], except that the angle of the BTZ description is not periodic. This angle corresponds to the spatial direction  $x$  of the 1+1 dimensional CFT and it becomes periodic if we compactify the theory<sup>11</sup> [49,50,51]<sup>12</sup>. All calculations done for the 1D+5D system [39,53,54] are evidence for this conjecture. In all these cases [54] the nontrivial part of the greybody factors comes from the  $AdS$  part of the spacetime. Indeed, it was noticed in [55] that the greybody factors for the BTZ black hole were the same as the ones for the five-dimensional black hole in the dilute gas approximation. The dilute gas condition  $r_0, r_n \ll r_1 r_5$  [53] is automatically satisfied in the limit (4.1) for finite temperature configurations (and finite chemical potential for the momentum along  $\hat{x}$ ). It was also noticed that the equations have an  $SL(2, R) \times SL(2, R)$  symmetry [56], these are the isometries of  $AdS_3$ , and part of the conformal symmetry of the 1+1 dimensional field theory. It would be interesting to understand what is the gravitational counterpart of the full conformal symmetry group in 1+1 dimensions.

## 5. Theories with $4 \rightarrow 8$ supersymmetries

The theories of this type will be related to black strings in five dimensions and Reissner-Nordström black holes in four dimensions. This part will be more sketchy, since there are several details of the conformal field theories involved which I do not completely understand, most notably the dependence on the various moduli of the compactification.

### 5.1. Black string in five dimensions

One can think about this case as arising from M-theory on  $M^6$  where  $M^6$  is a CY manifold,  $K3 \times T^2$  or  $T^6$ . We wrap fivebranes on a four-cycle  $P_4 = p^A \alpha_A$  in  $M^6$  with nonzero triple self intersection number, see [57]. We are left with a one dimensional object in five spacetime dimensions. Now we take the following limit

$$l_p \rightarrow 0 \quad (2\pi)^6 v \equiv V_6/l_p^6 = \text{fixed} \quad U^2 \equiv r/l_p^3 = \text{fixed} , \quad (5.1)$$

---

<sup>11</sup> I thank G. Horowitz for many discussions on this correspondence and for pointing out ref. [49] to me. Some of the remarks below arose in conversations with him.

<sup>12</sup> The ideas in [49,50,51] could be used to show the relation between the  $AdS$  region and black holes in M-theory on a light like circle. However the statement in [49,50,51] that the  $AdS_3 \times S^3$  spacetime is U-dual to the full black hole solution (which is asymptotic to Minkowski space) should be taken with caution because in those cases the spacetime has identifications on circles that are becoming null. This changes dramatically the physics. For examples of these changes see [32,52].

where  $l_p$  is the eleven dimensional Planck length. In this limit the theory will reduce to a conformal field theory in two dimensions. It is a (0,4) CFT and it was discussed in some detail in a region of the moduli space in [57]. More generally we should think that the five dimensional theory has some strings characterized by charges  $p^A$ , forming a multiplet of the U-duality group and we are taking a configuration where the triple self intersection number  $p^3$  is nonzero (in the case  $M^6 = T^6$ ,  $p^3 \equiv D \equiv D_{ABC}p^A p^B p^C$  is the cubic  $E_6$  invariant).

We now take the corresponding limit of the black hole solution. We will just present the near horizon geometry, obtained after taking the limit. Near the horizon all the vector moduli are at their fixed point values [58]. So the near horizon geometry can be calculated by considering the solution with constant moduli. We get

$$ds^2 = l_p^2 \left[ \frac{U^2 v^{1/3}}{D^{1/3}} (-dt^2 + dx^2) + \frac{D^{2/3}}{v^{2/3}} \left( 4 \frac{dU^2}{U^2} + d\Omega_2^2 \right) \right]. \quad (5.2)$$

In this limit  $M^6$  has its vector moduli equal to their fixed point values which depend only on the charge while its hyper moduli are what they were at infinity. The overall size of  $M^6$  in Planck units is a hypermultiplet, so it remains constant as we take the limit (5.1). We get a product of three dimensional  $AdS_3$  spacetime with a two-sphere,  $AdS_3 \times S^2$ . Defining the five dimensional Planck length by  $l_{5p}^3 = l_p^3/v$  we find that the “radii” of the two sphere and the  $AdS_3$  are  $R_{sph} = R_{AdS}/2 = l_{5p} D^{1/3}$ . In this case the superconformal group contains as a bosonic subgroup  $SO(2, 2) \times SO(3)$ . So the R-symmetries are just  $SU(2)_R$ , associated to the 4 rightmoving supersymmetries.

In this case we conjecture that this (0,4) conformal field theory is dual, for large  $p^A$ , to M-theory on  $AdS_3 \times S^2 \times M_p^6$ . The hypermultiplet moduli of  $M_p^6$  are the same as the ones entering the definition of the (0,4) theory. The vector moduli depend only on the charges and their values are those that the black string has at the horizon. A necessary condition for this conjecture to work is that the (0,4) theory should be independent of the original values of the vector moduli (at least for large  $p$ ). It is not clear to me whether this is true.

Using this conjecture we would get for large  $N$  a compactification of  $M$  theory which has five extended dimensions.

## 5.2. Extremal 3+1 dimensional Reissner-Nordström

This section is more sketchy and contains an unresolved puzzle, so the reader will not miss much if he skips it.

We start with IIB string theory compactified on  $M^6$ , where  $M^6$  is a Calabi-Yau manifold or  $K3 \times T^2$  or  $T^6$ . We consider a configuration of  $D3$  branes that leads to a black hole with nonzero horizon area. Consider the limit

$$\alpha' \rightarrow 0 \quad (2\pi)^6 v \equiv \frac{V_6}{\alpha'^3} = \text{fixed} \quad U \equiv \frac{r}{\alpha'} = \text{fixed}. \quad (5.3)$$

The string coupling is arbitrary. In this limit the system reduces to quantum mechanics on the moduli space of the three-brane configuration.

Taking the limit (5.3) of the supergravity solution we find

$$ds^2 = \alpha' \left[ \frac{U^2}{g_4^2 N^2} dt^2 + g_4^2 N^2 \frac{dU^2}{U^2} + g_4^2 N^2 d\Omega_2^2 \right] \quad (5.4)$$

where  $N$  is proportional to the number of  $D3$  branes. We find a two dimensional  $AdS_2$  space times a two-sphere, both with the same radius  $R = l_{4p}N$ , where  $l_{4p}^2 = g^2 \alpha'/v$ . The bosonic symmetries of  $AdS_2 \times S^2$  are  $SO(2, 1) \times SO(3)$ . This superconformal symmetry seems related to the symmetries of the *chiral* conformal field theory that was proposed in [59] to describe the Reissner-Nordström black holes. Here we find a puzzle, since in the limit (5.3) we got a quantum mechanical system and not a 1+1 dimensional conformal field theory. In the limit (5.3) the energy gap (mentioned in [60,59]) becomes very large<sup>13</sup>. So it looks like taking a large  $N$  limit at the same time will be crucial in this case. These problems might be related to the large ground state entropy of the system.

If this is understood it might lead to a proposal for a non perturbative definition of M/string theory (as a large  $N$  limit) when there are four non-compact dimensions.

It is interesting to consider the motion of probes on the  $AdS_2$  background. This corresponds to going into the ‘‘Coulomb’’ branch of the quantum mechanics. Dimensional analysis says that the action has the form (2.7) with  $p = 0$ . Expanding  $f$  to first order we find  $S \sim \int dt \frac{\dot{U}^2}{U^3} \sim \int dt v^2/r^3$ , which is the dependence on  $r$  that we expect from supergravity when we are close to the horizon. A similar analysis for Reissner-Nordström black holes in five dimensions would give a term proportional to  $1/r^4$  [17]. It will be interesting to check the coefficient (note that this is the *only* term allowed by the symmetries, as opposed to [17]).

---

<sup>13</sup> I thank A. Strominger for pointing this out to me.

## 6. Discussion, relation to matrix theory

By deriving various field theories from string theory and considering their large  $N$  limit we have shown that they contain in their Hilbert space excitations describing supergravity on various spacetimes. We further conjectured that the field theories are dual to the full quantum M/string theory on various spacetimes. In principle, we can use this duality to give a definition of M/string theory on flat spacetime as (a region of) the large  $N$  limit of the field theories. Notice that this is a non-perturbative proposal for defining such theories, since the corresponding field theories can, *in principle*, be defined non-perturbatively. We are only scratching the surface and there are many things to be worked out. In [61] it has been proposed that the large  $N$  limit of D0 brane quantum mechanics would describe eleven dimensional M-theory. The large  $N$  limits discussed above, also provide a definition of M-theory. An obvious difference with the matrix model of [61] is that here  $N$  is not interpreted as the momentum along a compact direction. In our case,  $N$  is related to the curvature and the size of the space where the theory is defined. In both cases, in the large  $N$  limit we expect to get flat, non-compact spaces. The matrix model [61] gives us a prescription to build asymptotic states, we have not shown here how to construct graviton states, and this is a very interesting problem. On the other hand, with the present proposal it is more clear that we recover supergravity in the large  $N$  limit.

This approach leads to proposals involving five (and maybe in some future four) non-compact dimensions. The five dimensional proposal involves considering the 1+1 dimensional field theory associated to a black string in five dimensions. These theories need to be studied in much more detail than we have done here.

It seems that this correspondence between the large  $N$  limit of field theories and supergravity can be extended to non-conformal field theories. An example was considered in [1], where the theory of NS fivebranes was studied in the  $g \rightarrow 0$  limit. A natural interpretation for the throat region is that it is a region in the Hilbert space of a six dimensional “string” theory<sup>14</sup>. And the fact that contains gravity in the large  $N$  limit is just a common feature of the large  $N$  limit of various field theories. The large  $N$  master field seems to be the anti-deSitter supergravity solutions [17].

When we study non extremal black holes in  $AdS$  spacetimes we are no longer restricted to low energies, as we were in the discussion in higher dimensions [44,54]. The

---

<sup>14</sup> This possibility was also raised by [62], though it is a bit disturbing to find a constant energy flux to the UV (that is how we are interpreting the radial dimension).

restriction came from matching the *AdS* region to the Minkowski region. So the five dimensional results [53,54] can be used to describe arbitrary non-extremal black holes in three dimensional Anti-deSitter spacetimes. This might lead us to understand better where the degrees of freedom of black holes really are, as well as the meaning of the region behind the horizon. The question of the boundary conditions is very interesting and the conformal field theories should provide us with some definite boundary conditions and will probably explain us how to interpret physically spacetimes with horizons. It would be interesting to find the connection with the description of 2+1 dimensional black holes proposed by Carlip [63].

In [8,13] super-singleton representations of *AdS* were studied and it was proposed that they would describe the dynamics of a brane “at the end of the world”. It was also found that in maximally supersymmetric cases it reduces to a free field [8]. It is tempting therefore to identify the singleton with the center of mass degree of freedom of the branes [6,13]. A recent paper suggested that super-singletons would describe all the dynamics of *AdS* [51]. The claim of the present paper is that all the dynamics of *AdS* reduces to previously known conformal field theories.

It seems natural to study conformal field theories in Euclidean space and relate them to deSitter spacetimes.

Also it would be nice if these results could be extended to four-dimensional gauge theories with less supersymmetry.

### Acknowledgments

I thank specially G. Horowitz and A. Strominger for many discussions. I also thank R. Gopakumar, R. Kallosh, A. Polyakov, C. Vafa and E. Witten for discussions at various stages in this project. My apologies to everybody I did not cite in the previous version of this paper. I thank the authors of [64] for pointing out a sign error.

This work was supported in part by DOE grant DE-FG02-96ER40559.

## 7. Appendix

$D = p + 2$ -dimensional anti-deSitter spacetimes can be obtained by taking the hyperboloid

$$-X_{-1}^2 - X_0^2 + X_1^2 + \cdots + X_p^2 + X_{p+1}^2 = -R^2 , \quad (7.1)$$

embedded in a flat  $D+1$  dimensional spacetime with the metric  $\eta = \text{Diag}(-1, -1, 1, \dots, 1)$ . We will call  $R$  the “radius” of *AdS* spacetime. The symmetry group  $SO(2, D - 1) =$

$SO(2, p+1)$  is obvious in this description. In order to make contact with the previously presented form of the metric let us define the coordinates

$$\begin{aligned} U &= (X_{-1} + X_{p+1}) \\ x_\alpha &= \frac{X_\alpha R}{U} \quad \alpha = 0, 1, \dots, p \\ V &= (X_{-1} - X_{p+1}) = \frac{x^2 U}{R^2} + \frac{R^2}{U} . \end{aligned} \tag{7.2}$$

The induced metric on the hyperboloid (7.1) becomes

$$ds^2 = \frac{U^2}{R^2} dx^2 + R^2 \frac{dU^2}{U^2} . \tag{7.3}$$

This is the form of the metric used in the text. We could also define  $\tilde{U} = U/R^2$  so that metric (7.3) has an overall factor of  $R^2$ , making it clear that  $R$  is the overall scale of the metric. The region outside the horizon corresponds to  $U > 0$ , which is only a part of (7.1). It would be interesting to understand what the other regions in the *AdS* spacetime correspond to. For further discussion see [65].

## References

- [1] J. Maldacena and A. Strominger, hep-th/9710014.
- [2] N. Seiberg, Phys. Lett. **B408** (1997) 98, hep-th/9705221
- [3] R. Haag, J. Lopuszanski and M. Sohnius, Nucl. Phys. **B88** (1975) 257.
- [4] W. Nahm, Nucl. Phys. **B135** (1978) 149.
- [5] G. Gibbons, Nucl. Phys. **B207**, (1982) 337; R. Kallosh and A. Peet, Phys. Rev. **D46** (1992) 5223, hep-th/9209116; S. Ferrara, G. Gibbons, R. Kallosh, Nucl. Phys. **B500** (1997) 75, hep-th/9702103.
- [6] G. Gibbons and P. Townsend, Phys. Rev. Lett. **71** (1993) 5223, hep-th/9307049.
- [7] Look for “gauged” supergravities in *Supergravities in Diverse Dimensions* , Vol. 1 and 2, A. Salam and E. Sezgin, (1989), North-Holland.
- [8] C. Frondal, Phys. Rev. **D26** (82) 1988; D. Freedman and H. Nicolai, Nucl. Phys. **B237** (84) 342; K. Pilch, P. van Nieuwenhuizen and P. Townsend, Nucl. Phys. **B242** (84) 377; M. Günaydin, P. van Nieuwenhuizen and N. Warner, Nucl. Phys. **B255** (85) 63; M. Günaydin and N. Warner, Nucl. Phys. **B272** (86) 99; M. Günaydin, N. Nilsson, G. Sierra and P. Townsend, Phys. Lett. **B176** (86) 45; E. Bergshoeff, A. Salam, E. Sezgin and Y. Tanii, Phys. Lett. **205B** (1988) 237; Nucl. Phys. **D305** (1988) 496; E. Bergshoeff, M. Duff, C. Pope and E. Sezgin, Phys. Lett. **B224** (1989) 71;
- [9] M. Günaydin and N. Marcus, Class. Quant. Grav. **2** (1985) L11; Class. Quant. Grav. **2** (1985) L19; H. Kim, L. Romans and P. van Nieuwenhuizen, Phys. Lett. **143B** (1984) 103; M. Günaydin, L. Romans and N. Warner, Phys. Lett. **154B** (1985) 268; Phys. Lett. **164B** (1985) 309; Nucl. Phys. **B272** (1986) 598
- [10] M. Blencowe and M. Duff, Phys. Lett. **203B** (1988) 229; Nucl. Phys. **B310** (1988) 387.
- [11] H. Nicolai, E. Sezgin and Y. Tanii, Nucl. Phys. **B305** (1988) 483.
- [12] M. Duff, G. Gibbons and P. Townsend, hep-th/9405124.
- [13] P. Claus, R. Kallosh and A. van Proeyen, hep-th/9711161.
- [14] G. Horowitz and A. Strominger, Nucl. Phys. **B360** (1991) 197.
- [15] G. Gibbons, G. Horowitz and P. Townsend, hep-th/9410073.
- [16] L. Susskind, hep-th/9611164; O. Ganor, S. Ramgoolam and W. Taylor IV, Nucl. Phys. **B492** (1997) 191, hep-th/9611202.
- [17] M. Douglas, J. Polchinski and A. Strominger, hep-th/9703031.
- [18] M. Aganagic, C. Popescu and J. Schwarz, Nucl. Phys. **B495** (1997) 99, hep-th/9612080.
- [19] M. Dine and N. Seiberg, Phys. Lett. **B409** (1997) 239, hep-th/9705057.
- [20] K. Becker, M. Becker, J. Polchinski and A. Tseytlin, Phys. Rev. **D56** (1997) 3174, hep-th/9706072.
- [21] A. Polyakov, hep-th/9711002.

- [22] T. Banks, W. Fishler, I. Klebanov and L. Susskind, hep-th/9709091.
- [23] S. Gubser, I. Klebanov and A. Peet, Phys. Rev. **D54** (1996) 3915, hep-th/9602135; A. Strominger, unpublised.
- [24] I. Klebanov, Nucl. Phys. **B496** (1997) 231, hep-th/9702076.
- [25] S. Gubser, I. Klebanov and A. Tseytlin, Nucl. Phys. **B499** (1997) 217, hep-th/9703040.
- [26] S. Gubser and I. Klebanov, hep-th/9708005.
- [27] C. Vafa, Nucl. Phys. **B469** (1996) 403, hep-th/9602022.
- [28] I. Klebanov and A. Tseytlin, Nucl. Phys. **B475** (1996) 164, hep-th/9604089.
- [29] A. Strominger, Phys. Lett. **B383** (1996) 44, hep-th/9512059.
- [30] E. Witten, hep-th/9507121, N. Seiberg and E. Witten, Nucl. Phys. **B 471** (1996) 121, hep-th/9603003
- [31] J. Maldacena, Proceedings of Strings'97, hep-th/9709099.
- [32] N. Seiberg, Phys. Rev. Lett. **79** (1997) 3577, hep-th/9710009.
- [33] A. Sen, hep-th/9709220.
- [34] N. Seiberg, hep-th/9705117.
- [35] S. Sethi and L. Susskind, Phys. Lett. **B400** (1997) 265, hep-th/9702101.
- [36] T. Banks and N. Seiberg, Nucl. Phys. **B497** (1997) 41, hep-th/9702187.
- [37] T. Banks, W. Fishler, N. Seiberg and L. Susskind, Phys. Lett. **B408** (1997) 111, hep-th/9705190.
- [38] I. Bars, Phys. Rev. **D55** (1997) 2373, hep-th/9607112.
- [39] A. Strominger and C. Vafa, Phys. Lett. **B379** (1996) 99, hep-th/9601029.
- [40] G. Horowitz, J. Maldacena and A. Strominger, Phys. Lett. **B383** (1996) 151, hep-th/9603109.
- [41] L. Andrianopoli, R. D'Auria and S. Ferrara, Int. J. Mod .Phys **A12** (1997) 3759, hep-th/9612105.
- [42] S. Ferrara, R. Kallosh and A. Strominger, Phys. Rev. **D52** (1995) 5412, hep-th/9508072; S. Ferrara and R. Kallosh, Phys. Rev. **D54** (1996) 1514, hep-th/9602136; Phys.Rev. **D54** (1996) 1525, hep-th/960309.
- [43] J.C. Breckenridge, R.C. Myers, A.W. Peet and C. Vafa, Phys. Lett. **B391** (1997) 93, hep-th/9602065.
- [44] J. Maldacena, Phys. Rev. **D55** (1997) 7645, hep-th/9611125.
- [45] D. Diaconescu and N. Seiberg, JHEP07(1997)001, hep-th/9707158.
- [46] O. Aharony, M. Berkooz, S. Kachru, N. Seiberg and E. Silverstein, hep-th/9707079.
- [47] E. Witten, hep-th/9707093.
- [48] Bañados, Teitelboim and Zanelli, Phys. Rev. Lett. **69** (1992) 1849, hep-th/9204099.
- [49] S. Hyun, hep-th/9704005.
- [50] H. Boonstra, B. Peeters and K. Skenderis, hep-th/9706192
- [51] K. Sfetsos and K. Skenderis, hep-th/9711138.
- [52] S. Hellerman and J. Polchinski, hep-th/9711037.

- [53] G. Horowitz and A. Strominger, Phys. Rev. Lett. **77** (1996) 2368, hep-th/9602051.
- [54] A. Dhar, G. Mandal and S. Wadia Phys. Lett. **B388** (1996) 51, hep-th/9605234; S. Das and S. Mathur, Nucl. Phys. **B478** (1996) 561, hep-th/9606185; Nucl.Phys. **B482** (1996) 153, hep-th/9607149; J. Maldacena and A. Strominger, Phys. Rev. **D55** (1997) 861, hep-th/9609026; S. Gubser, I. Klebanov, Nucl. Phys. **B482** (1996) 173, hep-th/9608108; C. Callan, Jr., S. Gubser, I. Klebanov and A. Tseytlin, Nucl. Phys. **B489** (1997) 65, hep-th/9610172; I. Klebanov and M. Krasnitz Phys. Rev. **D55** (1997) 3250, hep-th/9612051; I. Klebanov, A. Rajaraman and A. Tseytlin Nucl. Phys. **B503** (1997) 157, hep-th/9704112; S. Gubser, hep-th/9706100; H. Lee, Y. Myung and J. Kim, hep-th/9708099; K. Hosomichi, hep-th/9711072.
- [55] D. Birmingham, I. Sachs and S. Sen, hep-th/9707188.
- [56] M. Cvetic and F. Larsen, Phys. Rev. **D56** (1997) 4994, hep-th/9705192; hep-th/9706071.
- [57] J. Maldacena, A. Strominger and E. Witten, hep-th/9711053.
- [58] A. Chamseddine, S. Ferrara, G. Gibbons and R Kallosh, Phys. Rev. **D55** (1997) 3647, hep-th/9610155.
- [59] J. Maldacena and A. Strominger, Phys. Rev. **D56** (1997) 4975, hep-th/9702015.
- [60] J. Maldacena and L. Susskind, Nucl .Phys. **B475** (1996) 679, hep-th/9604042.
- [61] T. Banks, W. Fischler, S. Shenker and L. Susskind, hep-th/9610043.
- [62] O. Aharony, S. Kachru, N. Seiberg and E. Silverstein, private communication.
- [63] S. Carlip, Phys. Rev. **D51** (1995) 632, gr-qc/9409052.
- [64] R. Kallosh, J. Kumar and A. Rajaraman, hep-th/9712073.
- [65] S. Hawking and J. Ellis, *The large scale structure of spacetime*, Cambrige Univ. Press (1973), and references therein.

February 1, 2008

SLAC-PUB-7769  
SU-ITP-98/13

## The Hierarchy Problem and New Dimensions at a Millimeter

Nima Arkani-Hamed\*, Savas Dimopoulos\*\* and Gia Dvali†

\* SLAC, Stanford University, Stanford, California 94309, USA

\*\* Physics Department, Stanford University, Stanford, CA 94305, USA

† ICTP, Trieste, 34100, Italy

### Abstract

We propose a new framework for solving the hierarchy problem which does not rely on either supersymmetry or technicolor. In this framework, the gravitational and gauge interactions become united at the weak scale, which we take as the only fundamental short distance scale in nature. The observed weakness of gravity on distances  $\gtrsim 1$  mm is due to the existence of  $n \geq 2$  new compact spatial dimensions large compared to the weak scale. The Planck scale  $M_{Pl} \sim G_N^{-1/2}$  is not a fundamental scale; its enormity is simply a consequence of the large size of the new dimensions. While gravitons can freely propagate in the new dimensions, at sub-weak energies the Standard Model (SM) fields must be localized to a 4-dimensional manifold of weak scale “thickness” in the extra dimensions. This picture leads to a number of striking signals for accelerator and laboratory experiments. For the case of  $n = 2$  new dimensions, planned sub-millimeter measurements of gravity may observe the transition from  $1/r^2 \rightarrow 1/r^4$  Newtonian gravitation. For any number of new dimensions, the LHC and NLC could observe strong quantum gravitational interactions. Furthermore, SM particles can be kicked off our 4 dimensional manifold into the new dimensions, carrying away energy, and leading to an abrupt decrease in events with high transverse momentum  $p_T \gtrsim$  TeV. For certain compact manifolds, such particles will keep circling in the extra dimensions, periodically returning, colliding with and depositing energy to our four dimensional vacuum with frequencies of  $\sim 10^{12}$  Hz or larger. As a concrete illustration, we construct a model with SM fields localised on the 4-dimensional throat of a vortex in 6 dimensions, with a Pati-Salam gauge symmetry  $SU(4) \times SU(2) \times SU(2)$  in the bulk.

# 1 Introduction

There are at least two seemingly fundamental energy scales in nature, the electroweak scale  $m_{EW} \sim 10^3$  GeV and the Planck scale  $M_{Pl} = G_N^{-1/2} \sim 10^{18}$  GeV, where gravity becomes as strong as the gauge interactions. Over the last two decades, explaining the smallness and radiative stability of the hierarchy  $m_{EW}/M_{Pl} \sim 10^{-17}$  has been one of the greatest driving forces behind the construction of theories beyond the Standard Model (SM). While many different specific proposals for weak and Planck scale physics have been made, there is a commonly held picture of the basic structure of physics beyond the SM. A new effective field theory (e.g. a softly broken supersymmetric theory or technicolor) is revealed at the weak scale, stabilizing and perhaps explaining the origin of the hierarchy. On the other hand, the physics responsible for making a sensible quantum theory of gravity is revealed only at the Planck scale. The desert between the weak and Planck scales could itself be populated with towers of new effective field theories which can play a number of roles, such as triggering dynamical symmetry breakings or explaining the pattern of fermion masses and mixings.

In this picture, the experimental investigation of weak scale energies is quite exciting, as it is guaranteed to reveal the true mechanism of electroweak symmetry breaking and stabilization of the hierarchy. One can also hope that a detailed measurement of low energy parameters can give valuable clues to the structure of effective field theories at higher energies, perhaps even approaching the Planck scale. Nevertheless, it is fair to say that in this paradigm, the thorough exploration of the weak scale will never give a direct experimental handle on strong gravitational physics.

It is remarkable that such rich theoretical structures have been built on the assumption of the existence of two disparate fundamental energy scales,  $m_{EW}$  and  $M_{Pl}$ . However, there is an important difference between these scales. While electroweak interactions have been probed at distances  $\sim m_{EW}^{-1}$ , gravitational forces have not remotely been probed at distances  $\sim M_{Pl}^{-1}$ : gravity has only been accurately measured in the  $\sim 1\text{cm}$  range. Our interpretation of  $M_{Pl}$  as a fundamental energy scale (where gravitational interactions become strong) is then based on the assumption that gravity is unmodified over the 33 orders of magnitude between where it is measured at  $\sim 1\text{ cm}$  down to the Planck length  $\sim 10^{-33}\text{ cm}$ . Given the crucial way in which the fundamental role attributed to  $M_{Pl}$  affects our current thinking, it is worthwhile questioning this extrapolation and seeking new alternatives to the standard picture of physics beyond the SM.

In fact, given that the fundamental nature of the weak scale is an experimental certainty, we wish to take the philosophy that  $m_{EW}$  is the only fundamental short distance scale in nature, even setting the scale for the strength of the gravitational interaction. In this approach, the usual prob-

lem with the radiative stability of the weak scale is trivially resolved: the ultraviolet cutoff of the theory is  $m_{EW}$ . How can the usual  $(1/M_{Pl})$  strength of gravitation arise in such a picture? A very simple idea is to suppose that there are  $n$  extra compact spatial dimensions of radius  $\sim R$ . The Planck scale  $M_{Pl(4+n)}$  of this  $(4+n)$  dimensional theory is taken to be  $\sim m_{EW}$  according to our philosophy. Two test masses of mass  $m_1, m_2$  placed within a distance  $r \ll R$  will feel a gravitational potential dictated by Gauss's law in  $(4+n)$  dimensions

$$V(r) \sim \frac{m_1 m_2}{M_{Pl(4+n)}^{n+2}} \frac{1}{r^{n+1}}, \quad (r \ll R). \quad (1)$$

On the other hand, if the masses are placed at distances  $r \gg R$ , their gravitational flux lines can not continue to penetrate in the extra dimensions, and the usual  $1/r$  potential is obtained,

$$V(r) \sim \frac{m_1 m_2}{M_{Pl(4+n)}^{n+2}} \frac{1}{R^n r}, \quad (r \gg R) \quad (2)$$

so our effective 4 dimensional  $M_{Pl}$  is

$$M_{Pl}^2 \sim M_{Pl(4+n)}^{2+n} R^n. \quad (3)$$

Putting  $M_{Pl(4+n)} \sim m_{EW}$  and demanding that  $R$  be chosen to reproduce the observed  $M_{Pl}$  yields

$$R \sim 10^{\frac{30}{n}-17} \text{ cm} \times \left( \frac{1 \text{ TeV}}{m_{EW}} \right)^{1+\frac{2}{n}}. \quad (4)$$

For  $n = 1$ ,  $R \sim 10^{13}$  cm implying deviations from Newtonian gravity over solar system distances, so this case is empirically excluded. For all  $n \geq 2$ , however, the modification of gravity only becomes noticeable at distances smaller than those currently probed by experiment. The case  $n = 2$  ( $R \sim 100 \mu\text{m} - 1 \text{ mm}$ ) is particularly exciting, since new experiments will be performed in the very near future, looking for deviations from gravity in precisely this range of distances [11].

While gravity has not been probed at distances smaller than a millimeter, the SM gauge forces have certainly been accurately measured at weak scale distances. Therefore, the SM particles cannot freely propagate in the extra  $n$  dimension, but must be localized to a 4 dimensional submanifold. Since we assume that  $m_{EW}$  is the only short-distance scale in the theory, our 4-dimensional world should have a "thickness"  $\sim m_{EW}^{-1}$  in the extra  $n$  dimensions. The only fields propagating in the  $(4+n)$  dimensional bulk are the  $(4+n)$  dimensional graviton, with couplings suppressed by the  $(4+n)$  dimensional Planck mass  $\sim m_{EW}$ .

As within any extension of the standard model at the weak scale, some mechanism is needed in the theory above  $m_{EW}$  to forbid dangerous higher dimension operators (suppressed only by  $m_{EW}$ ) which lead to proton decay, neutral meson mixing etc. In our case, the theory above  $m_{EW}$  is unknown, being whatever gives a sensible quantum theory of gravity in  $(4+n)$  dimensions! We therefore simply assume that these dangerous operators are not induced. Any extension of the SM at the weak scale must also not give dangerously large corrections to precision electroweak observables. Again, there could be unknown contributions from the physics above  $m_{EW}$ . However, at least the purely gravitational corrections do not introduce any new electroweak breakings beyond the  $W, Z$  masses, and therefore should decouple as loop factor  $\times(m_{W,Z}/m_{EW})^2$ , which is already quite small even for  $m_{EW} \sim 1$  TeV.

Summarizing the framework, we are imagining that the space-time is  $R^4 \times M_n$  for  $n \geq 2$ , where  $M_n$  is an  $n$  dimensional compact manifold of volume  $R^n$ , with  $R$  given by eq. (4). The  $(4+n)$  dimensional Planck mass is  $\sim m_{EW}$ , the only short-distance scale in the theory. Therefore the gravitational force becomes comparable to the gauge forces at the weak scale. The usual 4 dimensional  $M_{Pl}$  is not a fundamental scale at all, rather, the effective 4 dimensional gravity is weakly coupled due to the large size  $R$  of the extra dimensions relative to the weak scale. While the graviton is free to propagate in all  $(4+n)$  dimensions, the SM fields must be localized on a 4-dimensional submanifold of thickness  $m_{EW}^{-1}$  in the extra  $n$  dimensions.

Of course, the non-trivial task in any explicit realization of this framework is localization of the SM fields. A number of ideas for such localizations have been proposed in the literature, both in the context of trapping zero modes on topological defects[7] and within string theory. In section 3, we will construct models of the first type, in which there are two extra dimensions and, given a dynamical assumption, the SM fields are localized within the throat of a weak scale vortex in the 6 dimensional theory. We want to stress, however, that this particular construction must be viewed at best as an “existence proof” and there certainly are other possible ways for realizing our proposal, without affecting its most important consequences.

It is interesting that in our framework supersymmetry is no longer needed from the low energy point of view for stabilizing the hierarchy, however, it may still be crucial for the self-consistency of the theory of quantum gravity above the  $m_{EW}$  scale; indeed, the theory above  $m_{EW}$  may be a superstring theory.

Independently of any specific realization, there are a number of dramatic experimental consequences of our framework. First, as already mentioned, gravity becomes comparable in strength to the gauge interactions at energies  $m_{EW} \sim \text{TeV}$ . The LHC and NLC would then not only probe the mechanism of electroweak symmetry breaking, they would probe the true quantum theory

of gravity!

Second, for the case of 2 extra dimensions, the gravitational force law should change from  $1/r^2$  to  $1/r^4$  on distances  $\sim 100\mu\text{m}$ -1 mm, and this deviation could be observed in the next few years by the new experiments measuring gravity at sub-millimeter distances[11].

Third, since the SM fields are only localized within  $m_{EW}^{-1}$  in the extra  $n$  dimensions, in sufficiently hard collisions of energy  $E_{esc} \gtrsim m_{EW}$ , they can acquire momentum in the extra dimensions and escape from our 4-d world, carrying away energy.\* In fact, for energies above the threshold  $E_{esc}$ , escape into the extra dimensions is enormously favored by phase space. This implies a sharp upper limit to the transverse momentum which can be seen in 4 dimensions at  $p_T = E_{esc}$ , which may be seen at the LHC or NLC if the beam energies are high enough to yield collisions with c.o.m. energies greater than  $E_{esc}$ .

Notice that while energy can be lost into the extra dimensions, electric charge (or any other unbroken gauge charge) can not be lost. This is because the massless photon is localized in our Universe and an isolated charge can not exist in the region where electric field can not penetrate, so charges can not freely escape into the bulk. In light of this fact, let us examine the fate of a charged particle kicked into the extra dimensions in more detail. On very general grounds (which we will discuss in more detail in section 3), the photon (or any other massless gauge field) can be localized in our Universe, provided it can only propagate in the bulk in the form of a massive state with mass  $\sim m_{EW}$ ,  $m_{EW}^{-1}$  setting the penetration depth of the electric flux lines into the extra dimensions. In order for the localized photon to be massless it is necessary that the gauge symmetry be unbroken at least within a distance  $\gg m_{EW}^{-1}$  from our four-dimensional surface (otherwise the photon will get mass through the “charge screening”, see section 3). As long as this condition is satisfied, the four-dimensional observer will see an unbroken gauge symmetry with the right 4-d Coulomb law. Now, consider a particle with nonzero charge (or any other unbroken gauge quantum number) kicked into the extra dimensions. Due to the conservation of flux, an electric flux tube of the width  $m_{EW}^{-1}$  must be stretched between the escaping particle and our Universe. Such a string has a tension  $\sim m_{EW}^2$  per unit length. Depending on the energy available in the collision, the charged particle will be either be pulled back to our Universe, or the flux tube will break into pieces with

---

\*Usually in theories with extra compact dimensions of size  $R$ , states with momentum in the compact dimensions are interpreted from the 4-dimensional point of view particles of mass  $1/R$ , but still localized in the 4-d world. This is because the at the energies required to excite these particles, their wavelength and the size of the compact dimension are comparable. In our case the situation is completely different: the particles which can acquire momentum in the extra dimensions have TeV energies, and therefore have wavelengths much smaller than the size of the extra dimensions. Thus, they simply escape into the extra dimensions.

opposite charges at their ends. In either case, charge is conserved in the 4-dimensional world, although energy may be lost in the form of neutral particles propagating in the bulk. Similar conclusions can be reached by considering a soft photon emission process[8].

Once the particles escape into the extra dimensions, they may or may not return to the 4-dimensional world, depending on the shape and/or the topology of the  $n$  dimensional compact manifold  $M_n$ . In the most interesting case, the particles orbit around the extra dimensions, periodically returning, colliding with and depositing energy to our 4 dimensional space with frequency  $R^{-1} \sim 10^{27-30/n}$  Hz. This will lead to continuous “fireworks”, which in the case of  $n = 2$  can give rise to  $\sim$  mm displaced vertices.

## 2 Phenomenological and Astrophysical Constraints

In our framework physics below a TeV is very simple: It consists of the Standard Model together with a graviton propagating in  $4+n$  dimensions. Equivalently—in four dimensional language—our theory consists of the Standard model together with the graviton and all its Kaluza-Klein (KK) excitations recurring once every  $1/R$ , per extra dimension  $n$ . We shall refer to all of them collectively as the “gravitons”, independent of their mass. Since each graviton couples with normal gravitational strength  $\sim 1/M_{Pl}$  to matter, its effect on particle physics and astrophysical processes is negligible. Nevertheless, since the multiplicity of gravitons beneath any relevant energy scale  $E$  is  $(ER)^n$  can be large, the *combined* effect of all the gravitons is not always negligible and may lead to observable effects and constraints. In this section we will very roughly estimate the most stringent of these constraints, mainly to show that our framework is not grossly excluded by current lab and astrophysical bounds. Clearly, a much more detailed study must be done to more precisely determine the constraints on  $n$  and  $m_{EW}$  in our framework.

Consider any physical process involving the emission of a graviton. The amplitude of this process is proportional to  $1/M_{Pl}$  and the rate to  $1/M_{Pl}^2$ . Consequently, the total combined rate for emitting any one of the available gravitons is

$$\sim \frac{1}{M_{Pl}^2} (\Delta E R)^n \quad (5)$$

where  $\Delta E$  is the energy available to the graviton and the last term counts the KK gravitons’ multiplicity for  $n$  extra dimensions. Using eq (3) we can rewrite this as

$$\sim \frac{\Delta E^n}{m_{EW}^{2+n}} \quad (6)$$

Note that the same result can be seen from the  $4 + n$  dimensional point of view. The  $m_{EW}$  suppressions of the couplings of the  $4 + n$  dimensional graviton are determined by expanding  $g_{AB} = \eta_{AB} + h_{AB}/\sqrt{m_{EW}^{2+n}}$ , where  $h_{AB}$  is the canonically normalised graviton in  $4 + n$  dimensions. Squaring this amplitude to obtain the rate yields precisely the  $m_{EW}$  dependence found above. As a result, the branching ratio for emitting a graviton in any process is

$$\sim (\Delta E/m_{EW})^{2+n} \quad (7)$$

The experimentally most exciting (and most dangerous) case has  $m_{EW} \sim$  TeV and  $n = 2$ . Of course, we must assume that weak-scale suppressed operators giving proton decay, large  $K-\bar{K}$  mixing etc. are forbidden. Of the remaining lab constraints, the ones involving the largest energy transfers  $\Delta E$  (such as  $\Upsilon$  and  $Z$  decays) are most constrained. The branching ratio for graviton emission in  $\Upsilon$  decays is unobservable  $\sim 10^{-8}$ . For  $Z \rightarrow X + \text{graviton}$  the branching ratio goes up to  $\sim 10^{-5}$ . Absence of such decay modes puts the strongest laboratory constraints to the scale  $m_{EW}$  and/or  $n$ . Nevertheless, they are easy to satisfy, in part because of their sensitivity to small changes in the value of  $m_{EW}$ . Production of gravitons in very high energy collisions will give the same characteristic signatures as the missing energy searches, except for one difference: the missing energy is now being carried by massless particles.

Next we consider astrophysical constraints. The gravitons are similar to goldstone bosons, axions and neutrinos in at least one respect. They can carry away bulk energy from a star and accelerate its cooling dynamics. For this reason their properties are constrained by the sun, red giants and SN 1987A. The simplest way to estimate these constraints is to translate from the known limits on goldstone particles. The dictionary that allows us to do that follows from eq(6) :

$$1/F^2 \longleftrightarrow \Delta E^n/m_{EW}^{2+n} \quad (8)$$

relating the emission rate of goldstones and gravitons. Here  $F$  is the goldstone boson's decay constant. For the sun the available energy  $\Delta E$  is only a keV. Therefore, even for the maximally dangerous case  $m_{EW} = 1$  TeV and  $n = 2$ , the effective  $F$  is  $10^{12}$  GeV, large enough to be totally safe for the sun; the largest  $F$  that is probed by the sun is  $\sim 10^7$  GeV.

For red giants the available energy is  $\sim 100$  keV and the effective  $F \sim 10^{10}$  GeV. This value is an order of magnitude higher than the lower limit from red giants. Finally we consider the supernova 1987A. There, the maximum available energy per particle is presumed to be between 20 and 70 MeV . Choosing the more favorable 20 MeV we find an effective  $F \sim 10^8$  GeV, which is smaller than the lower limit of  $10^{10}$  GeV claimed from SN 1987A. Therefore, the astrophysical theory of SN 1987 A places an interesting constraint on the fundamental scale  $m_{EW}$  or/and the number of extra dimensions

$n$ . The constraint is easily satisfied if  $n > 2$  or if  $m_{EW} > 10$  TeV. Of course, when the number of dimensions gets large enough so that  $1/R \gtrsim 100$  MeV, (corresponding to  $n \gtrsim 7$ ), none of the astrophysical bounds apply, since all the relevant temperatures would be too low to produce even the lowest  $KK$  excitation of the graviton.

Finally, although accelerators have not achieved collisions in the TeV energy range where the most exotic aspects of the extra dimensions are revealed, one may wonder whether very high energy cosmic rays of energies  $\sim 10^{15} - 10^{19}$  eV (which in colliding with protons correspond to c.o.m. energies  $\sim 1\text{-}100$  TeV) have already probed such physics. However, the cosmic rays are smoothly accelerated to their high energies without any “hard” interactions, and they have dominantly soft QCD interactions with the protons they collide with. Therefore, there are no significant constraints from very high energy cosmic ray physics on our framework.

Having outlined our general ideas, some dramatic experimental consequences and being reassured that existing data do not significantly constrain the framework, we turn to constructing an explicit model realising our picture, with SM fields localised on the four-dimensional throat of a vortex in 6 dimensions.

### 3 Construction of a Realistic Model.

In this section we construct a realistic model incorporating the ideas of this paper. As stressed in the introduction, this should be viewed as an example or an “existence proof”, since similar constructions are possible in the context of field theory as well as string theory. In particular one can change the structure and dimensionality of the manifold, the localization mechanism, the gauge group and the particle content of the theory without affecting the key ideas of our paper. Furthermore, many of the phenomenological consequences are robust and do not depend on such details.

The space time is 6-dimensional with a signature  $g_{AB} = (-1, 1, 1, 1, 1, 1)$ . The two extra dimensions are compactified on a manifold with a radius  $R \sim 1\text{mm}$ . We will discuss two possible topologies: a two-sphere and a two-torus with the zero inner radius. In both cases the key point is that the observable particles (quarks, leptons, Higgs and gauge bosons) are localized inside a small region of weak-scale size equal to the inverse cutoff length  $\sim \Lambda^{-1}$  and can penetrate in the bulk only in form of the heavy modes of mass  $\sim \Lambda$ . Thus for the energies  $< \Lambda$  ordinary matter gets confined to a four-dimensional hypersurface, our universe. The transverse  $x_5, x_6$  dimensions can be probed only through the gravitational force, which is the only long-range interaction in the bulk.

There are several ways to localize the Standard Model particles in our

four-dimensional space-time. Here we consider the possibility that localization is dynamical and the ordinary particles are “zero modes” trapped in the core of a four-dimensional vortex. This topological defect, in its ground state, is independent of four coordinates ( $x_\mu$ ) and thus carves-out the four-dimensional hypersurface which constitutes our universe.

Consider first  $x_5, x_6$  to be compactified on a two-sphere. Define a six-dimensional scalar field  $\Phi(x_A)$  transforming under some  $U(1)_V$  symmetry. We assume that  $\Phi$  gets a nonzero VEV  $\sim \Lambda$  and breaks  $U(1)_V$  spontaneously. The vortex configuration is independent of the four coordinates  $x_\mu$  and can be set up through winding the phase by  $2\pi$  around the equator of the sphere:

$$\Phi = \phi_{bulk} e^{i\theta} \quad (9)$$

where  $2\pi > \theta > 0$  is an azimuthal angle on the sphere and  $\phi_{bulk}$  is the constant expectation value that minimizes a potential energy (modulo the small curvature corrections). Such a configuration obviously implies two zeros of the absolute value  $\Phi$  on the both sides of the equator, which can be placed at the north and the south poles respectively. These zeros represent the vortex–anti-vortex pair of characteristic thickness  $\sim \Lambda^{-1}$ . Since this size is much smaller than the separation length  $\sim 1\text{mm}$ , vortex can be approximated by the Nielsen-Olesen solution [3]

$$\Phi = f(r) e^{i\theta}, \quad f(0) = 0, \quad f(r)|_{r>>\Lambda^{-1}} \rightarrow \phi_{bulk} \quad (10)$$

where  $0 < r < 2\pi R$  is a radial coordinate on the sphere, and an anti-vortex corresponds to the change  $\theta \rightarrow -\theta, r \rightarrow 2\pi R - r$ . If  $U(1)_V$  is gauged the magnetic flux will be entering the south pole and coming out from the north one.

Since any closed loop on a two-sphere can be smoothly deformed to a point, vortices on a sphere are not truly stable, and can annihilate with anti-vortices if they encounter one another. However the vortex anti-vortex pair are separated by a millimeter, which is  $10^{16}$  times their size; they therefore are, for all practical purposes, stable. In addition there can be other mechanisms of stabilization if other forces are involved (e.g. repelling charges or currents flowing along the strings, etc.).

Alternatively, compactification on a torus can lead to a truly stable vortex. This is because a torus contains many non-contractible loops, and the phase of  $\Phi$  winding along such a loop is topologically stable. Such a configuration is obtained from the previously discussed two-sphere by identifying its poles with a single point and subsequently removing this point from the manifold. This manifold is then equivalent to a two-torus with zero inner radius; it can carry topological charge and accommodate a single vortex on it. The magnetic flux goes through the point that was removed from the manifold. An observer looking at the south pole will see the vortex with

incoming flux. If he travels towards the north pole along the meridian he will arrive to the same vortex, since the poles have been identified, but will see it as an anti-vortex since he will now be looking at the flux up-side down.

Next, we come to the localization of the standard model particles on a vortex. We discuss the localization of different spins separately.

### 3.1 Localization of Fermions

Fermions can be trapped on the vortex as “zero modes” [1] because of the index theorem [2]. Consider a pair of six-dimensional left-handed Weyl spinors  $\Gamma_7\psi, \xi = \psi, \xi$ . These, in terms of the four-dimensional chiral Weyl spinors can be written as

$$\psi = (\psi_L, \psi_R), \quad \xi = (\xi_L, \xi_R) \quad (11)$$

Assume now that this pair is getting a mass from coupling to the vortex field:

$$h\Phi\psi\xi + \text{h.c.}, \quad (12)$$

where  $h \sim \Lambda^{-1}$  has dimensions of inverse mass. The six-dimensional Dirac equation in the vortex background is:

$$\Gamma_A \partial^A \psi^+ = h\phi_{bulk} e^{i\theta} \xi, \quad (13)$$

and similarly for  $\xi^+$ . To look for solutions with localized massless fermions we separate variables through the angle-independent anzatz  $\psi = \psi(x_\mu)\beta(r)$  and  $\xi = \xi(x_\mu)\beta(r)$ , where  $\mu = 1,..4$ ,  $\beta(r)$  is a radial scalar function in the 2 dimensional compact space of  $x_5$  and  $x_6$ . To be zero modes of the four-dimensional Dirac operator, the spinors  $\psi(x_\mu)$  and  $\xi(x_\mu)$  must satisfy

$$\Gamma_5 e^{i\theta(-i\Gamma_5\Gamma_6)} \psi^+ \partial_r \beta(r) = h\phi_{bulk} e^{i\theta} \xi, \quad (14)$$

and similarly for  $\xi^+$ . Since  $\psi^+$  and  $\xi$  must be eigenvalues of the  $(-i\Gamma_5\Gamma_6)$  operator, they automatically have definite four-dimensional chirality (say left for the vortex and right for the anti-vortex). In this case the normalizable wavefunction has the localized radial dependence  $\beta(r) = e^{-h \int_0^r f(r') dr'}$ . Thus the vortex supports a single four-dimensional massless chiral mode which can be

$$\psi_L + \xi_R^+ \quad (15)$$

In general, as a consequence of the index theorem[1][2], every pair of six-dimensional chiral fermions getting mass from the vortex field, deposits a single zero mode of definite four-dimensional chirality. Thus through the couplings to the vortex field one can reproduce the whole set of the four-dimensional chiral fermions – quarks and leptons– localized on the submanifold. These localized modes can get nonzero masses through the usual Higgs

mechanism. Let  $\psi$  and  $\psi'$  be the six-dimensional chiral spinors (from different pairs) that deposit two different zero modes on the vortex. These zero modes can get masses through the couplings to a scalar field  $H$

$$H\psi\psi', \quad (16)$$

provided it has a nonzero expectation value in the core of the vortex but vanishes in the bulk. The index is unaffected by the existence of such a scalar since it has a zero VEV outside the core.

### 3.2 Localization of Higgs Scalars

Now let us consider how the Higgs fields with non-zero VEVs can be localized on the vortex. A massive scalar field can be easily localized provided it has a suitable sign coupling to the vortex field in the potential

$$h'|\Phi|^2|H|^2 \quad (17)$$

If  $h' > 0$ , this contribution is positive in the bulk and zero in the core. Thus  $H$  will see the defect as an attractive potential, which for a certain range of parameters can lead to a bound-state solution. We will treat  $H$  as the six-dimensional progenitor of the Weinberg-Salam Higgs particle. Then the physically most important case is when  $H$  develops a non-zero expectation value in the the defect. This is easily possible, provided an effective mass<sup>2</sup> of the Higgs becomes negative in the throat of the vortex. The simplest prototype potential of this sort is:

$$(h'|\Phi|^2 - m^2)HH^+ + c(HH^+)^2, \quad \text{with } m^2, h', c > 0, \quad (18)$$

where  $H$  is a six-dimensional scalar field. We will assume that  $h'\Phi_{bulk}^2 - m^2 > 0$  and thus  $H$  is zero in the bulk. However, it can develop VEV in the vortex core. This does not require any fine tuning and can be seen by examining the stability of the trivial solution with respect to the small excitations  $H(x_5, x_6)e^{-i\omega t}$  in the vortex background. The analysis is similar to the one of the superconducting cosmic string [4]. The linearized equation for small excitations is the two dimensional Schrodinger equation

$$-\partial_{5,6}^2 H + [h'f - m^2]H = \omega^2 H \quad (19)$$

which certainly has a normalizable boundstate solution with  $\omega^2 < 0$  in a range of parameters. Thus  $H$  becomes tachionic in the core marking the instability of the trivial solution  $H = 0$ . As a result  $H$  develops an expectation value in the throat of the vortex which decays exponentially for large  $r$ . Since  $H$  is a Higgs doublet there are the three massless Goldstone modes localized on the vortex . These get eaten up by  $W$  and  $Z$  bosons through the usual Higgs

effect. There is also a localized massive mode, an ordinary Higgs scalar, which corresponds to the small vibrations of the expectation value in the core  $H(0) \rightarrow H(0) + h(x_\mu)$ . Such a vibration propagates in four-dimensions as the ordinary massive scalar Higgs.

### 3.3 Localization of Gauge Fields

There are several possible mechanisms for localizing gauge fields on a vortex (or on any other topological defect)<sup>†</sup> through the coupling to the vortex scalar. In general, a particle localized in such a way will not be massless, unless there is a special reason such as the index theorem for fermions and the Goldstone theorem or supersymmetry for bosons. Here, we propose to localize massless gauge fields by generalizing the four-dimensional confinement mechanism of ref[5] (see also [8]). Before discussing how this mechanism is generalized to our case, it is instructive to understand why the simplest approach fails. Consider the  $U(1)_{EM}$  electromagnetism in the presence of a thin vortex along the z-axis. Let  $\phi^-$  be a charged scalar field that develops an expectation value and breaks  $U(1)_{EM}$  spontaneously through the Higgs mechanism. Now introducing a cross-coupling with the vortex field

$$(-a|\Phi|^2 + m^2)|\phi^-|^2 + b|\phi^-|^2, \quad a, b, m^2 > 0 \quad (20)$$

and appropriately adjusting parameters (no fine tuning), we can force  $\phi^-$  to vanish in the vortex<sup>‡</sup>; as a consequence, the tree-level mass of the six-dimensional electric field coupled to it will also vanish. Unfortunately, the four-dimensional photon trapped in this way does not remain massless. It has a massive dispersion relation due to charge screening. This can be understood as follows: Since the charged field  $\phi^-$  condenses in the vacuum, the Universe is superconducting everywhere except in the interior of the thin vortex. Two test charges placed at different points  $x_\mu$  and  $x'_\mu$  in the vortex will not interact by Coulomb's law; their electric field polarizes the surrounding medium, and the field lines end on the superconductor. As a result, the electric flux along the vortex dies-off exponentially with (longitudinal) distance within a characteristic length given by the width of the vortex.

It is clear that for the localized gauge field to be massless the surrounding medium which repels the electric field lines should not contain any charge condensate, otherwise all the field lines can be absorbed by the medium. To construct such an example, consider a thin planar-layer between two infinite superconductors. Two magnetic monopoles located inside the layer interact through a long range magnetic field. This is because the magnetic flux is

---

<sup>†</sup> An alternate way to localize massless gauge fields involves D-brane constructions[10]

<sup>‡</sup> An alternative possibility is that  $\phi^-$  is a vortex field itself, charged under the electromagnetism, just as in the Abrikosov vortices in superconductors. In this case  $\phi^-$  will vanish at the origin for topological reasons.

repelled (or “totally reflected”) from the superconductor, since it contains no magnetic charges on which the magnetic field lines can end. Consequently, the magnetic flux is entirely contained inside the layer and, as a result of flux conservation, the field lines spread according to Coulomb’s law.

In ref[5] a dual version of this mechanism –in which the superconductor is replaced by a confining medium with monopole condensation– was suggested as a way to obtain massless gauge bosons localized on a sub-manifold. Suppose that away from the vortex  $U(1)_{EM}$  becomes a part of a confining group which develops a mass gap  $\sim \Lambda$ . Then the electric flux lines will be repelled by monopole condensation in the dual Meisner effect; no images are created since there is no charged condensate in the medium.

It is not difficult to construct an explicit four-dimensional prototype model of this sort. It includes an  $SU(2)$  Yang-Mills theory with a scalar field  $\chi$  in the adjoint representation, plus a vortex field  $\Phi$ , which breaks some additional  $U(1)_V$  symmetry and forms the string (for the present discussion it is inessential whether  $U(1)_V$  is global or gauged). The Lagrangian has the form ( $SU(2)$  indices are suppressed)

$$\begin{aligned} \mathcal{L} = & -\frac{1}{4g^2} \text{Tr} G_{\mu\nu} G_{\mu\nu} + (D_\mu \chi)^2 - \lambda(\chi^2)^2 - \chi^2(h|\Phi|^2 - M^2) + \\ & + |\partial_\mu \Phi|^2 - \lambda'(|\Phi|^2 - \phi_{bulk}^2)^2, \end{aligned} \quad (21)$$

where  $G_{\mu\nu}$  is the “gluon” field strength tensor, and  $h, \phi_{bulk}^2, M^2, \lambda, \lambda'$  are the positive parameters and we assume  $h\phi_{bulk}^2 > M^2$ . In a certain range of parameters the absolute minimum of the theory is achieved for  $\chi = 0$ . In this vacuum,  $\Phi$  develops the VEV  $\langle \Phi \rangle = \phi_{bulk}$  and forms the vortex. Although  $\chi$  is zero in the vacuum, it can acquire an expectation value inside the vortex, where its mass<sup>2</sup> becomes negative, just like in the example with the Higgs doublet considered above. In this case  $SU(2)$  is broken to  $U(1)_{EM}$  on the string, but is restored outside. Inside the string, two out of the three gluons acquire large masses of order of  $M$ . The third gluon becomes a photon. Two degrees of freedom in the  $\chi$  field are eaten up by the Higgs mechanism, the remaining degree of freedom is neutral. The massless degree of freedom in the effective  $1 + 1$  dimensional theory on the string is a photon. It is massless, since the  $U(1)_{EM}$  gauge symmetry is unbroken everywhere. On the other hand outside the vortex the photon becomes a member of the nonabelian gauge theory, which confines and develops a mass gap. Thus the photon can only escape from the string in the form of a composite heavy “glueball” with a mass of the order of  $\Lambda$  which we take to be the UV cutoff  $\sim m_{EW}$ . This guarantees that at low energies the massless photon will be trapped on the string. The theory inside the string is in the abelian  $1 + 1$  dimensional “Coulomb” phase.

How is this mechanism generalized to our six-dimensional case? Of course, we do not know how confinement works in a higher dimensional theory. Nev-

ertheless, we believe and will postulate that the higher dimensional theory in the bulk will posses a mass gap  $\sim \Lambda$  provided that:

- 1) Outside the vortex the standard model gauge group, in particular electromagnetism and strong interactions, are extended into a larger nonabelian gauge theory.
- 2) There is no light (with a mass below the cut-off scale  $\Lambda$ ) matter in the bulk enforced by general principles, such as Goldstone's theorem.
- 3) The tree-level gauge coupling blows up away from the vortex [8].

The latter condition can be satisfied e.g. if the value of the gauge coupling is set by an expectation value of the higgs field (or any function of it) which vanishes away from the vortex. For instance, in the previous four-dimensional toy model such a coupling is

$$\Lambda^{-2} \text{Tr} \chi^2 \text{Tr} G_{\mu\nu} G^{\mu\nu} \quad (22)$$

In summary, we presented one possible way for localizing particles in our four-dimensional space-time. There are other possibilities within both field and string theory –such as D-brane constructions– for accomplishing the same goal.

### 3.4 A Realistic Theory

In this section we assemble the above ingredients to construct a prototype model incorporating the ideas of this paper. We embed the Standard Model in the Pati-Salam group  $G = SU(4) \otimes SU(2)_R \otimes SU(2)_L$  which is the unbroken gauge group in the bulk. In addition, we introduce a  $U(1)_V$  factor and a singlet scalar field  $\Phi$  charged under it.  $\Phi$  develops an expectation value and forms a vortex of thickness  $\sim \Lambda^{-1}$  in the compact 2-D submanifold spanned by  $x_5, x_6$ . The interior of the vortex is our 4-dimensional space-time with all the light matter confined to it. The only light particle propagating in the bulk is the six-dimensional graviton.

The gauge group is spontaneously broken to  $SU(3) \otimes U(1)_{EM}$  inside the vortex, by a set of six-dimensional scalar fields  $\chi = (15.1.1)$ ,  $\chi' = (4.2.1)$  and  $H = (1.2.2)$  which develop nonzero VEVs only in the core of the vortex due to their interactions with the  $\Phi$  field. We assume a soft hierarchy  $\chi' \sim \chi \sim \Lambda \sim 10H \sim m_{EW}$ . The crucial assumption is that in the bulk the gauge group is strongly coupled and develops a mass gap of the order of the cut-off. This, together with the fact that  $SU(3) \otimes U(1)_{EM}$  is unbroken everywhere guarantees that the gluons and the photon are massless and trapped in our four-dimensional manifold.  $W^\pm$  and a  $Z$  bosons are localized as massive states.

The matter fermions are assumed to originate from the following six-dimensional chiral spinors per generation:

$$Q = (4, 1, 2), \bar{Q} = (\bar{4}, 1, 2), Q_c = (\bar{4}, 2, 1), \bar{Q}_c = (4, 2, 1), \quad (23)$$

which get their bulk masses through the coupling to the vortex field

$$h\Phi Q\bar{Q} + h'\Phi^* Q_c \bar{Q}_c \quad (24)$$

(where  $h$  and  $h'$  are parameters of the inverse cut-off size). The index theorem ensures that each pair deposits a single chiral zero mode which can be chosen as  $Q_L + \bar{Q}_R^+$  and  $Q_{cL}^+ + \bar{Q}_{cR}$ . These states get their masses through the couplings to the Higgs doublet field which condenses in the core of the vortex

$$gHQQ_c + \bar{g}H\bar{Q}\bar{Q}_c \quad (25)$$

To avoid unacceptable flavor violations, the couplings in eqs(24) should be flavor-universal. This can be guaranteed by some flavor symmetry. Flavor violations must come from the ordinary Yukawa couplings (see eq(25) to be under control.

The theory presented here has rich and exotic phenomenology, thanks to the existence of the extra dimensions. At energies above the cutoff of a  $\Lambda \sim$  TeV there is a plethora of particles which can quite freely migrate and allow us to look into the extra dimensions. Furthermore, naturalness requires that the migration into the extra dimensions cannot be postponed much beyond the TeV scale.

## 4 Summary and Outlook

The conventional paradigm for High Energy Physics –which dates back to at least 1974– postulates that there are two fundamental scales , the weak interaction and the Planck scale. The large disparity between these scales has been the major force driving most attempts to go beyond the Standard Model, such as supersymmetry and technicolor. In this paper we propose an alternate framework in which gravity and the gauge forces are united at the weak scale. As a consequence, gravity lives in more than four dimensions at macroscopic distances –leading to potentially measurable deviations from Newton’s inverse square law at sub-mm distances. The LHC and NLC are now even more interesting machines. In addition to their traditional role of probing the electroweak scale, they are quantum-gravity machines, which can also look into extra dimensions of space via exotic phenomena such as apparent violations of energy, sharp high- $p_T$  cutoffs and the disappearance and reappearance of particles from extra dimensions.

The framework that we are proposing changes the way we think about some fundamental issues in particle physics and cosmology. The first and most obvious change in particle physics occurs in our view of the hierarchy problem. Postulating that the cutoff is at the weak scale nullifies the usual argument about ultraviolet sensitivity, since the weak scale now becomes the ultraviolet! The new hierarchy that we now have to face, in the six

dimensional case, is that between the millimeter and the weak scales. This hierarchy is stable in the sense that small changes of parameters have small effects on the physics –so there is no fine tuning problem. There is also no issue of radiatively destabilizing the mm scale by physics at the weak cutoff. In this respect, our proposal shares the same “set it and forget it” philosophy of the original proposal supersymmetric standard model [12]. An important and favorable difference is that the mm scale is not a Lagrangean parameter that needs to be stabilized by a symmetry, such as supersymmetry. It is a parameter characterizing a solution, the size of the two extra dimensions. It is not uncommon to have solutions much larger than Lagrangean parameters; the world around us abounds with solutions that are much larger than the electron’s Compton-wavelength. A related secondary question is whether the magnitude of the mm scale may be calculated in a theory whose fundamental length is the weak scale. We have not addressed this question which is imbedded in the higher dimensional theory. It is amusing to note that if there are many new dimensions, their size –given by eq (4)– approaches the weak scale and there is no large hierarchy.

Finally we come to the early universe. The most solid aspect of early cosmology, namely primordial nucleosynthesis, remains intact in our framework. The reason is simple: The energy per particle during nucleosynthesis is at most a few MeV, too small to significantly excite gravitons. Furthermore, the horizon size is much larger than a mm so that the expansion of the universe is given by the usual 4-dimensional Robertson-Walker equations. Issues concerning very early cosmology, such as inflation and baryogenesis may change. This, however, is not necessary since there may be just enough space to accommodate weak-scale inflation and baryogenesis.

In summary, there are many new interesting issues that emerge in our framework. Our old ideas about unification, inflation, naturalness, the hierarchy problem and the need for supersymmetry are abandoned, together with the successful supersymmetric prediction of coupling constant unification [12]. Instead, we gain a fresh framework which allows us to look at old problems in new ways. Lagrangean parameters become parameters of solutions and the phenomena that await us at LHC, NLC and beyond are even more exciting and unforeseen.

**Acknowledgments:** We would like to thank I. Antoniadis, M.Dine, L. Dixon, N. Kaloper, A. Kapitulnik, A. Linde, M. Peskin, S. Thomas and R. Wagoner for tuseful discussions. G. Dvali would like to thank the Institute of Theoretical Physics of Stanford University for their hospitality. NAH is supported by the Department of Energy under contract DE-AC03-76SF00515. SD is supported by NSF grant PHY-9219345-004.

## References

- [1] R.Jackiw and P.Rossi, Nucl. Phys. B190, (1981) 681;
- [2] R.Jackiw and C.Rebbi, Phys. Rev.D13 (1976) 3398; E.Weinberg, Phys.Rev. D24 (1981) 2669.
- [3] H.B. Nielsen, P. Olesen, Nucl.Phys.B61 (1973) 45.
- [4] E. Witten, Nucl.Phys.B249 (1985) 557.
- [5] G. Dvali and M. Shifman, Phys.Lett.B396 (1997) 64.
- [6] L. Dixon, J.A. Harvey, C. Vafa and E. Witten, Nucl.Phys.B261 (1985) 678.
- [7] See, e.g., V.Rubakov and M.Shaposhnikov, Phys. Lett. 125B (1983) 136; G.Dvali and M.Shifman, Nucl.Phys.B504 (1997) 127.
- [8] See A.Barnaveli and O.Kancheli, Sov. J. Nucl. Phys. 51 (3), (1990).
- [9] For a review see, e.g. J. E. Kim, Phys. Rep. 150, (1987) 1; M.S.Turner, Phys.Rep.197 (1990) 67-97.
- [10] J. Polchinski , “TASI Lectures on D-Branes”, hep-th/9611050 .
- [11] J.C. Price, in proc. Int. Symp. on Experimental Gravitational Physics, ed. P.F. Michelson, Guangzhou, China (World Scientific, Singapore 1988); J.C.Price et. al., NSF proposal 1996; A.Kapitulnik and T. Kenny, NSF proposal, 1997.
- [12] S.Dimopoulos and H.Georgi, Nucl. Phys. B193,(1981), 150.





