Fundamentals
of
Quantum Mechanics

PRENTICE-HALL PHYSICS SERIES

Donald H. Menzel, *Editor*

Fundamentals
of
Quantum Mechanics

by

ENRICO PERSICO
DIRECTOR OF THE DEPARTMENT OF PHYSICS
LAVAL UNIVERSITY, QUEBEC

Translated and Edited
by

GEORGES M. TEMMER
RESEARCH ASSOCIATE, DEPARTMENT OF PHYSICS
UNIVERSITY OF ROCHESTER

PRENTICE-HALL, INC.
Englewood Cliffs, N. J.

First printing . . . . *October, 1950*
Second printing . . . . *May, 1954*
Third printing *January, 1957*

PRINTED IN THE UNITED STATES OF AMERICA
34270

Author's Preface

IN THIS BOOK are presented the fundamental laws and principal methods of quantum mechanics, together with some applications. For this reason only a few indicative notes will be found on the vast field of spectroscopy (with the exception of the theory of the hydrogen spectrum, which has a fundamental character). Furthermore, the theoretical interpretation of the periodic system of the elements, the statistical theories, the theory of the chemical bond, and problems of atomic collisions are not included.

Instead we have dwelled somewhat at length on the epistemological foundations of the new atomic mechanics. For anyone starting a study of this field, a major difficulty is the comprehension of its epistemological position, which is both unusual and profound. To help overcome this difficulty an exponent of the theory may choose various approaches. These range from the intuitive treatment (of necessity less rigorous, being based on analogies and qualitative discussions) to the strictly logical approach, which is bound to be abstract and formal. The former will be especially agreeable to mentalities of the "visual" type but will leave dissatisfied the critical mind with a logical inclination. The situation is reversed for the latter approach. In practice, it is convenient to strike a compromise in such a manner as to satisfy all types of readers. In this book, the author has attempted to resolve various exigencies by dividing the work into three parts of increasing orders of abstraction.

The first part serves to give a historical, overall view of the evolution of quantum mechanics. The approach here, as intuitive and elementary as is possible, describes some of the more important experiments upon which the theory is based, and gives a first glimpse of that theory.

The second part, of a more advanced nature, is preceded by a mathematical introduction. Here the principles of quantum mechanics in the particular form due to Schrödinger (*wave mechanics*) are established on the basis of the uncertainty principle. The most important problems involving a single particle are treated by this method. The old quantum theory of Bohr and Sommerfeld, which is still very useful from the didactic standpoint, is developed in the last chapter of the second part, rather than before wave mechanics as is customary. In fact, it appears that this arrangement, though not in historical order, is most logical—because the postulates of the Bohr-Sommerfeld theory do not seem to be *a priori* logically justifiable, though they may be obtained as a first approximation from the laws of wave mechanics.

Finally, in the third part, the principles of quantum mechanics in their most general form are presented (*transformation theory*). More advanced mathematical methods are used here (Hilbert space and matrices), which are briefly explained in the introductory chapter. However, in this last part also, recourse to the more abstract mathematical methods was avoided, but at the expense of maximum generality and rigor. Anyone desirous of a more profound exposition of quantum mechanics, from the point of view of its logical structure, is referred to the works of Neumann and Weyl cited in the Bibliography.

ENRICO PERSICO

Translator's Preface

THE PHYSICAL CONSTANTS occurring in this book have been replaced by their most recent values, obtained mainly from J. W. M. DuMond and E. R. Cohen, *Rev. Mod. Phys.* **20,** 82 (1948) and *ibid.* **21,** 651 (1949), and R. T. Birge, *Rev. Mod. Phys.* **13,** 233 (1941); the references pertaining to these constants have been brought up to date.

A number of supplementary notes and corrections supplied by Professor Persico were incorporated into the book. I have taken the liberty to insert additional footnotes and occasional sentences throughout the text whenever I thought it advisable in the interest of clarity; some changes were warranted by recent developments in the field of atomic physics. The list of books in the Bibliography has been enlarged by including a number of new references to books in the English language, and titles of books already listed have been supplemented by those of their English translations in cases where such translations exist. Otherwise, the original text has been left unaltered with regard to both content and style insofar as the difference in idiom would permit.

I am grateful to Professor Birge for help concerning the general constants, and to Odette Temmer for her aid in the preparation of the manuscript.

GEORGES M. TEMMER

Contents

Part I

HISTORICAL DEVELOPMENT AND EXPERIMENTAL BASES

Part II

WAVE MECHANICS OF A PARTICLE

Part III

GENERAL METHODS OF QUANTUM MECHANICS

Part I

HISTORICAL DEVELOPMENT AND EXPERIMENTAL BASES

CHAPTER 1

The Atomic Model

1. The concept of the atom in chemistry and physics. Although the hypothesis of a corpuscular structure of matter had been enunciated since earliest times, it became necessary for the interpretation of experimental facts only when the fundamental chemical laws of constant proportions and of multiple proportions were discovered. Indeed, these laws naturally lead one to consider every simple body to be made up of identical particles. These particles can combine in various ways, forming innumerable compounds, but they always occur as indivisible entities in all chemical phenomena—as *atoms*, in the etymological sense of the term. Chemistry permits us to determine the *ratios* of the weights of the various atoms with great accuracy, or to express all atomic weights in definite units in terms of one of them (notably of the oxygen atom). It does not, however, enable us to compare the weight of an atom with the weights of ordinary bodies; hence we cannot obtain atomic weights in fractions of a gram. Chemistry gives even less indication of the size and shape of atoms; it is content with the realization that atoms are so small as to escape direct observation. Therefore in order to understand the major part of chemical phenomena, it is sufficient to represent the atoms, for instance, as minute, hard, indivisible spheres (whose diameter is not of interest) possessing a particular kind of attractive force by means of which they can group themselves into *molecules*. The latter, when in turn assembled into large aggregates, constitute the bodies we see.

This oversimplified conception of the structure of matter is not only sufficient to account for chemical phenomena but also permits us to develop the theory of many physical phenomena. An example is the kinetic theory of gases, which obtains a large part of its results by considering the molecules as simple points or spheres, or aggregates of spheres. Furthermore, it is found that the mass m of the individual molecules enters into many formulas of the kinetic theory of gases, but that in the final expressions destined to be compared with experience, m appears *almost invariably* multiplied by the total number of molecules. Actually, then, only the total *mass* of the gas occurs; consequently, the major part of the theory is independent of the size attributed to the molecules, and similarly of their number. It is therefore not possible to draw any conclusions concerning these quantities. For this reason, up until fairly recently, the absolute size of molecules and atoms was ignored in the development of atomic theory.

Later on, however, phenomena were discovered in which the mass of a single molecule occurs directly, or at least *Avogadro's number* N, the number of molecules contained in one gram-molecular weight.[1] The Brownian movement is such a phenomenon. These phenomena have made possible the experimental determination of the number that, so to speak, represents the key for the passage from the world of ordinary bodies to the world of the atoms. It has even been possible to find Avogadro's number by about 15 independent methods. The fact that under the atomic hypothesis results have been obtained that agree within the limits of experimental error is the best proof of its validity.[2]

The dimensions of molecules or atoms enter into another group of phenomena (for example the viscosity of gases), and hence the latter have made it possible to determine these dimensions, at least as far as their order of magnitude is concerned. It was found that atoms have diameters of the order of 10^{-8} cm and that the distances at which atoms arrange themselves in forming molecules are of the same order, so that for molecules dimensions of the order of 10^{-8} cm

---

[1] As is known, one gram-molecular weight is the number of grams equal to the molecular weight. Therefore, if Avogadro's number N is known, the weight in grams of a single molecule is obtained by dividing the molecular weight by N. Similarly, the weight in grams of one atom is obtained by dividing the atomic weight by N.

[2] The most reliable value (1947) of Avogadro's number is $N = 6.0235 \times 10^{23}$, with an uncertainty of several units in the last significant figure [J. W. M. DuMond and E. R. Cohen, *Rev. Mod. Phys.* **21**, 651 (1949)]. From this value may be derived the unit adopted by the chemists, namely, $\frac{1}{16}$ of the atomic weight of oxygen (a mixture of various isotopes, cf. §4), equal to $1/N = 1.6602 \times 10^{-24}$ g; the weight of one hydrogen atom differs very slightly from this value (1.6732×10^{-24} g). In physics a slightly smaller unit is often used to express atomic weights, namely, $\frac{1}{16}$ of the weight of the most abundant isotope of oxygen. Multiplication by 1.00027 of weights expressed in chemical units converts them to the corresponding weights on the physical scale.

likewise result. Later on it will be seen how the modern theories furnish values for them with remarkable precision in the various cases.

However, the concept of atoms as simple and inert fragments of matter is incapable of explaining a large number of physical phenomena. Indeed, the emission and absorption of light, magnetic phenomena, X rays, radioactivity, phenomena that occur upon the passage of electricity through gases, electrolysis, the photoelectric effect, the thermionic effect, and others, all lead us to regard atoms as more or less complex mechanisms, composed of still smaller parts and containing positive and negative electric charges, normally in equal quantities. These electric charges always occur as multiples of an elementary charge e. Thus we are led to attribute a corpuscular structure to electricity as well as to matter, and to consider the electrical corpuscles as constituents of atoms. Atoms have thus lost their original significance of "indivisible" particles (indeed, diverse phenomena are known today, such as the rather common one of ionization, in which the atom is "divided"), and are instead considered to be mechanical systems, formed from several particles of different kinds (positive and negative electrons, protons, neutrons), with which we shall deal later on. Some of these particles are charged positively, others negatively, and still others are neutral; all are in motion with respect to each other under the action of mutual forces.

This complex concept of the atom as a structure containing electricity in motion, together with the electromagnetic nature of light, opens the possibility of understanding the interaction of matter and radiation as represented by such phenomena as emission and absorption, scattering, and resonance. Indeed, it is known from electromagnetic theory that the motion of electric charges can produce electromagnetic waves which, in turn, may set in motion electric charges upon which they are falling. Furthermore, since electric charges in motion generate a magnetic field, we may foresee the possibility of explaining the magnetic properties of matter in this way. It may even be observed that one of the first not entirely static conceptions of the atom (or molecule) was formulated by Ampère for the express purpose of reducing magnetic phenomena to electric currents circulating among the smallest constituents of matter.

Nevertheless, as we shall see shortly, these ample possibilities, which we can now perceive upon a superficial examination of the question, were realized only with considerable difficulty and led further than could have been suspected at first sight.

2. The electron. The negative electric particles, or *electrons*, have long been known to be the constituents of cathode rays as well as of beta rays from radioactive substances. Furthermore, they are emitted from incandescent metals (*thermionic effect*) and from many other substances when they are struck by light of short wavelength (*photoelectric effect*) or by X rays. The electrons obtained in these various ways are always identical in all their properties. By means of the deflection suffered in an electric and magnetic field, it has been possible to measure the ratio e/m of charge to mass, so that the value of m may be obtained from an independent measurement of e. The electronic charge e has been measured with great precision by C. T. R. Wilson and R. A. Millikan, by the "cloud-chamber method" and the "oil-drop method," respectively. The latter method permits the determination of the charge of the electrons taken one at a time, and from its results we can assert that the charges of all electrons are actually equal to each other, and that the value found is not just an average value.[3] The most reliable values of the charge (in absolute value) and mass of the electron are (1947):[4]

Charge: $e = 4.8029 \times 10^{-10}$ e.s.u. $= 1.602 \times 10^{-20}$ e.m.u.
Mass (for low velocities): $m_0 = 9.1055 \times 10^{-28}$ g.

The mass turns out to be about 1837 times smaller than the mass of the hydrogen atom. For high velocities, it is found that the ratio e/m suffers a decrease proportional to $\sqrt{1 - v^2/c^2}$, which is in agreement with the theory of relativity (if it is assumed, as seems reasonable, that the charge e remains unchanged), according to which the mass m should vary as

$$m = \frac{m_0}{\sqrt{1 - (v^2/c^2)}},$$

where m_0 is the value of the mass for low velocities.

[3] For a description of these experiments, see Nos. 27, 28, and 31 of the Bibliography.

[4] For a critical discussion of these and other physical constants, cf. R. T. Birge, *Rev. Mod. Phys.* **13**, 233 (1941) and *Reports on Progress in Physics*, Vol. VIII, p. 90, 1942; J. W. M. DuMond and E. R. Cohen, *Rev. Mod. Phys.* **20**, 82 (1948) and **21**, 651 (1949). We adopt DuMond and Cohen's values, unless otherwise stated.

It has been determined that in electrolysis a gram-molecular weight of monovalent ions transports a charge of 96,496 coulombs, equaling 2.8927×10^{14} e.s.u. Dividing this quantity by Avogadro's number (6.023×10^{23}), we obtain the charge of each monovalent ion, which coincides with the value recently reported for the electronic charge.[5] This agreement leads to the conclusion that the electric charge of monovalent ions is due to an electron being added to or removed from a neutral atom, and similarly that ions of higher valences are atoms having a certain number of electrons more or less than neutral atoms. In short, the electrons observed as free particles in cathode rays, in beta rays, and so on, are identified with those manifesting themselves in electrolytic phenomena.

Furthermore, atoms always contain electrons of the kind observed in the free state, as is proved by many facts which will be discussed presently. We shall confine these considerations to pointing out one of the proofs which historically has been of the greatest importance: in the Zeeman effect, the ratio e/m is determined experimentally for the electric particle to whose motion the emission of light is attributed, and a value is found which agrees with the one already determined for electrons by the deflection method. Thus we recognize the presence of electrons within atoms and their direct participation in the phenomena of light emission.

3. The positive particles. The Rutherford atom. Although the facts cited above, along with others, have for a long time definitely established the presence of negative electricity inside the atoms in the form of electrons, we have yet to see the form in which the positive electricity is found which is necessary to give the atom its characteristics of an electrically neutral system. In 1902, J. J. Thomson advanced the hypothesis that the atom is made up of a homogeneous sphere of positive electricity within which the electrons are immersed, like dust particles in a water drop. By the mutual action between positive and negative charges, the electrons would be attracted (as may easily be shown) toward the center of the sphere, with a force proportional to the distance, and would occupy positions of equilibrium under the action of that force and their mutual repulsions. However, this hypothesis was readily shown to be inadequate for the interpretation of the experimental facts.

---

[5] Here we have one of the best ways of determining N.

The most direct refutation of the Thomson hypothesis was made through a famous experiment performed by Lord Ernest Rutherford in 1911 on the passage of alpha particles through matter. These particles, which are known to be doubly ionized (that is, positively charged) helium atoms, emitted by radioactive substances with velocities of the order of 10^9 cm/sec, are able to traverse thin metallic foils by virtue of their small size and high velocity. When a narrow beam of alpha particles strikes a foil, it does not proceed in a straight line but is scattered into various directions, thereby proving that alpha particles are deviated upon traversing matter—some more, some less, and some even directly backwards. These deviations cannot be due to the collisions of alpha particles with electrons contained in the substance traversed, because the latter have a mass several thousand times smaller than that of the alpha particles and hence cannot deflect them appreciably. Consequently, these deflections must be due to the action of the remainder of the atom, that is, to the part which, in addition to the positive charge, contains almost the entire mass of the atom. If one accepts the Thomson model, it is easily seen that the repulsion exerted by the sphere of positive electricity upon the alpha particles increases at first upon approaching it, but when the particle penetrates into the interior of the sphere, the force *decreases* as the particle nears the center. The resulting trajectory will be curved in the manner shown in Fig. 1(a).

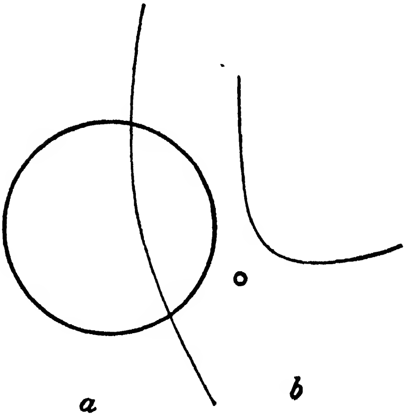

Fig. 1

The deviations obtained in this way are not very large and are insufficient to explain the experimental results. Therefore, Rutherford enunciated the hypothesis (which is universally accepted today) that the positive part of the atom consists of a *nucleus*, so small compared with atomic dimensions that one may consider it to be almost a point, in which nearly the entire mass of the atom is concentrated, not merely the positive charge. The repulsive force which the nucleus exerts upon the alpha particle will then be inversely proportional to the square of the distance and will increase indefinitely when the particle approaches the nucleus. Hence it is understandable that a considerably larger deviation must result than in the Thomson model [Fig. 1(b)]. A calculation of the angular distribution of the scattered beam of alpha particles yields a distribution in perfect agreement with the observed one (with the exception of those particles which pass very close to the nucleus, for which the nucleus may no longer be considered a point). Experience therefore confirms Rutherford's hypothesis of the "point nucleus." It is even possible, from the results of these experiments, to deduce the charge of the nucleus. In this manner the rather remarkable fact is discovered that the charge of the nucleus (except for the sign) is a multiple of the electronic charge equal to the *atomic number*, or order number in the periodic system of Mendelyeev, of the element considered.

On the basis of these results, Rutherford proposed a model of the atom which has been of fundamental importance in the development of theoretical physics. He considered the atom of atomic number Z to be constituted of a point *nucleus* in which almost the entire mass is concentrated, having a positive charge Ze (calling e the absolute value of the electronic charge) around which are found Z electrons, whose negative charges exactly compensate for the positive charge of the nucleus. In accordance with Coulomb's law, these electrons are attracted by the nucleus with a force inversely proportional to the square of the distance, similar to the force of gravity the sun exerts upon the planets. They will therefore move, according to Kepler's laws, in elliptical orbits (if we neglect the mutual repulsions between electrons); whereas the nucleus, because of its large mass, will remain practically at rest. The whole atom will thus represent a planetary system in miniature. The hydrogen atom, for which $Z = 1$, will have a nucleus of charge e and a single planetary electron. (This electron will describe an exactly elliptical orbit, since it lacks the perturbations of other electrons.) The helium atom will have a nucleus of charge $2e$ and two planetary electrons, and so on, up to curium ($Z = 96$), which will have a nuclear charge $96e$ and 96 electrons. (Elements with $Z = 43$, 61, 85, 87, 93, 94, 95, and 96 are unstable and do not occur in nature.)

According to this model, the quantity which is rather vaguely called the "diameter of the atom" (and which we have seen to be of the order of 10^{-8} cm) must be approximately identified with the maximum dimensions of the outer orbits. We can say that the atom very nearly occupies a sphere of that diameter; we understand, of course, that this sphere is not filled with solid substance but is almost entirely empty, and has particles traveling within it whose dimensions are very small compared with the distances between them.[6]

This model, which dominated atomic mechanics up until 1925, must not be interpreted as a faithful picture of intra-atomic reality but only as an approximation. Nevertheless, it retains great importance as a heuristic or didactic means, and it serves to furnish the basis for a convenient and expressive terminology, of which we shall make current use.

4. The structure of the nucleus. Isotopes. According to the Rutherford model, the valence and, in general, the chemical properties of an atom depend exclusively on the motion of the outermost electrons. The same can be said about the optical properties, such as, for example, the emission and absorption spectra. The innermost electrons intervene in determining the behavior of the atom with respect to X-ray spectra of emission and absorption. Finally, in atoms with radioactive properties, we must suppose that the latter properties reside within the nucleus, as has been confirmed by the fact that these properties are absolutely insensitive to chemical bonds, which alter the motion of the outermost electrons. They are also unaffected by X rays which act upon the ones closest to the nucleus, and in general by all physical agents known, excepting collisions with extremely penetrating particles or radiations.

This fact leads us to regard the nucleus not as an elementary particle such as the electron but as a generally complex system; this viewpoint is confirmed by various other circumstances, one of which is the existence of *isotopes.*

Although all elements can be considered to be simple bodies from the chemical standpoint, they may not be so regarded from the physical point of view. It turns out that elements may be further decomposed into as many as ten substances having identical chem-

---

[6] The radius of the nucleus is of the order of 10^{-12} cm; the radius of the electron is considered not to exceed 10^{-13} cm.

ical properties (with different atomic weights) and almost identical physical properties, with the exception of the possibility of radioactive properties, which are radically different. Each of these substances is called an *isotope.* The name, which means "same place" in Greek, is derived from the fact that in the periodic table, the various isotopes belonging to the same element must be thought of as being located in the same cell, since each cell corresponds to definite chemical properties.

Each isotope is made up of identical atoms, but the atoms of various isotopes differ in weight and other characteristics: thus it is the isotopes, not the elements, which correspond to separate atomic species. An element is generally constituted of a mixture of atoms of as many species as there are isotopes making up the element. The proportions of this mixture remain the same in all the compounds involving this element, as a consequence of the identity of the chemical properties of the isotopes belonging to the same elements. It turns out that these proportions are approximately the same no matter what the mineral from which the substance is obtained.

The existence of isotopes was first discovered through the radioactive elements. More recently, the work of Aston has shown that many of the nonradioactive substances, which were thought to be simple, are in reality mixtures of two, three, or more isotopes of different atomic weight. These mixtures behave like simple substances, since the various isotopes constituting them have almost perfectly identical physical and chemical properties. Hence the atomic weights as determined by chemists are really the result of an average of the atomic weights of the various isotopes. The latter have been determined with great precision, by physical methods which we shall not discuss here. The most important result of these investigations is that the atomic weights of the single isotopes are always *integers* to a good approximation. The nonintegral values occurring in the usual table of chemical elements derive from the fact that these weights do not refer to really simple substances but to mixtures, as has been pointed out. Thus, potassium, for instance, with atomic weight 39.1, is actually a mixture of two isotopes of atomic weights 39 and 41. The only physical properties by which two isotopes may appreciably differ are radioactive properties (both spontaneous and induced), and therefore the discovery and separation of isotopes has been much easier among naturally radioactive elements.

The interpretation of isotopes with the Rutherford model is as follows. The isotopes of the same element have the same atomic number, hence the same nuclear charge and the same number of planetary electrons, but they differ in the mass of the nucleus. Since the motion of the electrons is determined by electrical forces, and since the mass of the nucleus, which remains almost constant from one isotope to the next, exerts a very little influence upon the motion of the electrons, the latter move in almost the same manner in all isotopes; hence their identical physical and chemical properties. An exception is made by possible radioactive properties, which, as has been stated, do not depend on the planetary electrons but only on the structure of the nucleus, which is different in different isotopes.

The radioactive properties and their theoretical interpretation by means of appropriate hypotheses concerning the structure of nuclei constitute the subject matter of *nuclear physics.* This field, which received its theoretical basis only in recent years, has made enormous progress in an extremely short time, especially on the experimental side with the discovery of new disintegrations and of artificial radioactivity. During the Second World War, nuclear physics entered the field of practical applications with the invention of *nuclear piles* and of *atomic bombs*, which utilize the energy of atomic nuclei, commonly called "atomic energy." However, the physics of the nucleus is not within the scope of the present volume,[7] since in the questions of atomic physics with which we shall be concerned, the nucleus will always enter as a single particle of negligible dimensions. This nucleus will be characterized solely by the value of its mass and of its positive electric charge (or else by the atomic number). Nevertheless, attention should be called to the fact that this theory of nuclear structure (which is in a stage of active development) makes use of the same general principles of atomic mechanics which are treated in this volume.

We shall give here a list of the various kinds of elementary (or supposedly elementary) particles which have so far been discovered

---

[7] See F. Rasetti, *Elements of Nuclear Physics,* New York: Prentice-Hall, Inc., 1936; also No. 20b of the Bibliography.

experimentally. We shall give their masses expressed in atomic mass units ($\frac{1}{16}$ of the atomic mass of the most abundant isotope of oxygen), and the electric charge expressed in terms of the elementary charge $e = 4.8024 \times 10^{-10}$ e.s.u.

		Mass	*Charge*
Heavy particles	*proton*	1.007574	$+e$
	neutron	1.00893	0
Intermediate particles	*positive mesons*	~ 0.11; ~ 0.16	$+e$
	negative mesons	~ 0.11; ~ 0.16	$-e$
Light particles	*positive electron*	0.00054862	$+e$
	negative electron	0.00054862	$-e$

According to modern views, the nuclei of all atoms are aggregates of protons and neutrons (in particular, the nucleus of hydrogen consists of a proton only).

The mesons (positive and negative) are still relatively little known; they constitute a part of cosmic radiation, and have only recently (1948) been obtained artificially. For each sign of the charge, there exist at least two types of masses equal to about 200 and about 300 electron masses, respectively, or equaling 0.11 and 0.16 atomic mass units. These values, however, are known with considerably less accuracy than those of the other particles. The name "mesons" (or "mesotrons") derives from the fact that their mass is intermediate between that of the electron and that of the heavy particles.

The positive electrons (or *positrons*) and negative electrons (or *negatrons*) have the same mass and equal charges in absolute value. However, the symmetry between these two particles is not complete: although the negative electrons are found in all atoms (whose extranuclear part they constitute, as has been stated) and can be obtained very easily in the free state (beta rays, thermionic and photoelectric effects), the positive electrons have a very short life, having a tendency to unite with negative electrons (annihilation). Positrons are found in cosmic rays and are emitted by some artificially radioactive substances. Hence it is not surprising that the positron was discovered only in 1932, whereas the negative electron has been known since 1897.

The particles which we have listed are those whose existence has been confirmed experimentally; to these one must add some particles whose existence is considered plausible for theoretical reasons, although no direct experimental confirmation exists as yet. They are the *neutrino*, of negligible mass compared to the electron mass, and of zero charge; and the *neutretto*, or neutral meson, of mass intermediate between that of the electron and the proton, and of zero charge.

5. The radioactive displacement law. The identity of the atomic number with the nuclear charge is confirmed by the so-called *radioactive displacement law*. Indeed, it is found that when a radioactive substance emits negative beta rays, it is transformed into another having the same atomic weight but an atomic number greater by one unit (that is, it has been "displaced" one cell ahead in the periodic system). As a matter of fact, the nucleus, in losing one electron, does not vary appreciably in weight, whereas its positive charge increases by e. On the other hand, when a substance emits alpha rays, it is transformed into another having an atomic weight smaller by four units and an atomic number smaller by two units (that is, it has been displaced backward by two cells on the periodic chart). An alpha particle (which is known to be identical with a helium nucleus) has a weight of four and a charge $2e$; hence the weight of the nucleus which emitted it must diminish by four units, and the charge must decrease by $2e$.

In particular, if two beta emissions and one alpha emission follow each other in any order, the atomic weight should diminish by four but the nuclear charge should remain unchanged; we should thus end up with another isotope of the same element. This process actually occurs in certain cases. For example, radium D ($A = 210$, $Z = 82$), emitting a beta particle, is transformed into radium E ($A = 210$, $Z = 83$), which in turn emits another beta particle, thus becoming polonium ($A = 210$, $Z = 84$), which finally, by emitting an alpha particle, is transformed into radium G ($A = 206$, $Z = 82$). Now it is found that radium G has the same chemical properties as radium D; that is, they are both isotopes of the same element (lead).

A radioactive displacement law analogous to the one referring to negative beta rays also holds for the emission of positrons, which takes place in many artificially radioactive substances: in that case the isotope, without appreciably changing in atomic weight, is displaced backward by one cell; that is, its atomic number diminishes by one unit, in accordance with the loss of an elementary positive charge.

All these facts plainly show that all chemical and physical properties of an atom (except radioactive properties) which determine its position on the periodic table are themselves determined by the value of its nuclear charge, and that the "atomic number," originally defined as an order number on that table, can be identified with the number of elementary charges carried by the nucleus.

CHAPTER 2

Energy Quanta and Light Quanta

6. Inapplicability of classical laws to atoms. Although the Rutherford model of the atom is very attractive because of its simplicity, it leads to results in full disagreement with experience if we assume that its behavior is governed by the ordinary laws of mechanics and electricity.

For instance, according to the laws of electromagnetic theory, an electric charge radiates electromagnetic energy whenever it possesses a certain acceleration. Thus an electron which rotates around the nucleus, having a constant centripetal acceleration, should constantly radiate electromagnetic waves. Consequently, its energy should gradually diminish; this would lead to a gradual reduction of the dimensions of its orbit so that the electron would finally fall into the nucleus. Therefore the Rutherford atom could not have a permanent character; and its life could be calculated to be of the order of 10^{-8} sec.

Furthermore, the energy would be emitted in the form of radiation whose fundamental frequency would coincide with the frequency of the orbital motion of the electron. But since the latter frequency would be continually varying because of the shrinking of the orbit, the emitted light would have a changing frequency; therefore any body, containing countless atoms in all possible phases of their "lives," should emit radiation of all possible frequencies, or a continuous spectrum. However, it is known that gases emit line spectra of rigorously constant frequency.

When these difficulties presented themselves to the physicists, the latter had already been convinced for other reasons that classical mechanics and electromagnetic theory did not apply in the atomic domain, so that, rather than abandon the Rutherford model, they sought to find laws which would make it function in such a manner as to account for the experimental results. These laws were proposed for the first time by Niels Bohr in 1913, and subsequently stated in a more general form by A. Sommerfeld. However, prior to an exploration of these laws, we must rapidly outline those other fields in which contradictions with classical laws had already been encountered, and the attempts at the substitution of new laws. The first of these attempts, which opened the road to modern atomic physics, was that of the German physicist Max Planck, who in 1900 tried to give a theory of black-body radiation. An additional contribution concerning the nature of light rather than of matter was made by Einstein in 1905, introducing the concept of "light quanta" or "photons."

7. The spectrum of the black body. It is well known that incandescent bodies emit light; or, more generally, bodies emit radiation at any temperature, visible at higher temperatures but invisible for lower temperatures (infrared radiation). The type of radiation emitted depends, to a certain extent, on the nature of the body. However, if the latter is a "black body," that is, a body capable of absorbing all radiation it receives (such as lampblack, to a certain degree of approximation), one can show thermodynamically, and experiment confirms the result, that the radiation emitted at a given temperature is independent of the nature and of the shape of the body. When this radiation is analyzed with a spectroscope, it shows a continuous spectrum, whose intensity has been measured exactly at all points and at various temperatures. The intensity exhibits a maximum for a certain frequency ν_m and decreases toward zero for high as well as for low frequencies, as is indicated by the diagram in Fig. 2.

It is natural to suppose that the emitted radiation is due to the thermal agitation of the electric charges contained within matter; but if we try to express this idea quantitatively by applying the ordinary laws of mechanics and electromagnetic theory, we arrive at an absurd result. Instead of the experimental curve, we find an ascending parabolic curve which, near the origin, coincides with the experimental curve but then deviates from it in the manner shown in Fig. 2. The parabola is represented by the so-called Rayleigh-Jeans formula

$$I = \frac{2\pi}{c^2} kT\nu^2,$$

where k is Boltzmann's constant, c is the velocity of light, T is the absolute temperature, and $I(\nu)\,d\nu$ represents the energy radiated per unit area in the spectral band between ν and $\nu + d\nu$. According to this formula, the total energy radiated at any temperature different from absolute zero is infinite, as can be seen immediately by integrating with respect to ν.

Various attempts were made to escape this difficulty; but so long as use was made of classical mechanics and electromagnetic theory, they invariably led to the Rayleigh-Jeans formula, even when the particular mechanism of emission was modified in various ways. While making one of these attempts, toward the end of 1899, Planck[1] discovered at which point one had to break away from the classical treatment in order to reach a result conforming fully to experience.

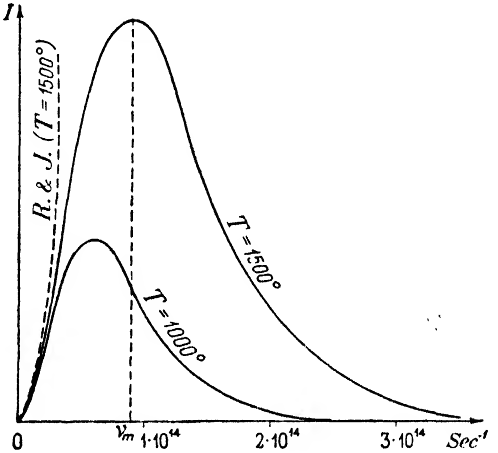

Fig. 2

He started out with the hypothesis that in atoms there are contained tiny *oscillators*[2] of all possible frequencies, in constant vibration (each with its proper frequency); these oscillators continually

---

[1] *Verh. d. D. Phys. Ges.* **2**, 237 (1900); *Ann. d. Physik* **4**, 553 (1901).

[2] An *oscillator* is a system formed by an electrically charged mass point (such as an electron) attracted to a fixed center with a force proportional to the distance. It is known from mechanics that such a point executes isochronous oscillations about the fixed center, with a frequency independent of amplitude and characteristic of the oscillator. However, Planck's reasoning does not necessarily require this model but remains valid under more general assumptions.

absorb and emit energy in the form of electromagnetic waves. Through an artifice of calculation, Planck attacked the problem by assuming that the energy of each oscillator varies not continuously but in steps of a certain magnitude ϵ, reserving the right of letting the latter tend toward zero later. But he found that if he permitted ϵ to go to zero, he would again be led to the Rayleigh-Jeans formula; on the other hand, when he let ϵ retain a finite value proportional to the proper frequency of the oscillator, by setting

$$\epsilon = h\nu \tag{1}$$

(h being constant), he arrived at the formula

$$I = \frac{2\pi\nu^2}{c^2}\frac{h\nu}{e^{h\nu/kT} - 1}, \tag{2}$$

which, for a fixed parameter T, represents a curve of the type shown in solid line in Fig. 2. Furthermore, upon assigning to h a convenient numerical value, which according to the latest determinations[3] is, in c.g.s. units,

$$h = 6.6237 \times 10^{-27},$$

formula (2) exactly represents the energy distribution in the spectrum of the black body for any value of the temperature T. Following this surprising result, Planck proposed the bold hypothesis that energy, like matter, possesses an atomic structure; that is, it occurs always in amounts that are multiples of an elementary quantity ϵ, which he called a *quantum*.[4] The latter must have, according to what has been said, a value proportional to the frequency with which the system under consideration is able to oscillate.

Thus there entered for the first time into science the concept of a discontinuity in the physical laws, in contrast to the ordinary mechanical and electromagnetic theories. The old theories were called *classical*, in order to distinguish them from the new ones, called *quantum theories*. Corresponding to this new concept, there entered into physics a new universal constant h, called *Planck's constant*, or *quantum of action*, whose importance was subsequently revealed to be very great in a multitude of phenomena.

The extreme hypothesis of an atomic structure for energy was

---

[3] Cf. J. W. M. DuMond and E. R. Cohen, *loc. cit.*

[4] Latin word used in German in a sense meaning "quantity," "dose."

soon reduced in scope by Planck himself, who showed that it is sufficient to assume a discontinuity only in the process of emission and not in the nature of the energy itself. Although, even in that form, the hypothesis was repugnant to the minds of many scientists, the success of its further developments in accounting for the specific heats of solids (theories of Einstein and Debye, 1907 and 1912) confirmed many people in their conviction that the hypothesis was a step in the right direction.

Today the formula of Planck can be justified by means independent of the oscillator hypothesis. In the case of the oscillator, the hypothesis of energy quanta appears as a particular consequence of rather general laws. These laws are based upon hypotheses which seem considerably more satisfactory and plausible than the original hypothesis of quanta, because they are devoid of contradictions both within themselves and with the laws which ordinary bodies obey; they also account in unique fashion for a very large number of phenomena.

8. The photoelectric effect and "light quanta." While in the field of the structure of matter, Planck's hypothesis gives rise to the idea that classical laws would not be applicable to the atom, Einstein pointed out a no less serious contradiction between certain experimental facts and the wave theory of light.

This contradiction, to which we shall turn in greater detail below, essentially consists in the following. It is known that if radiation of sufficiently high frequency (in general, ultraviolet light, X rays, or gamma rays) is made to fall upon a metallic surface, the latter emits electrons with a certain energy, which is without a doubt imparted to them by the incident radiation. The phenomenon is called the *photoelectric effect* in the case where the exciting radiation consists of visible or ultraviolet light. We shall, however, use the term also in the more general case. Now it is found that the energy which each single electron receives from the radiation (an energy which is partly used to tear the electron from the metal and partly remains with the electron as kinetic energy) is independent of the light intensity but depends solely upon its frequency; the intensity influences only the number of electrons emitted (which is proportional to the intensity) and not the energy with which each electron is emitted. Thus one is forced to abandon the most natural hypothesis: that the electrons are constrained to oscillate under the action of the electric field of the light and that, when their oscillations reach sufficiently large amplitudes, they are severed from the atom to which they belong and are ejected. However a more serious difficulty presents itself independently of any particular hypothesis concerning the mechanism of the photoelectric effect: when the intensity of the light is diminished, a point can be reached where the energy incident upon each atom during the entire experiment (calculated under the assumption that the energy falls uniformly upon the whole surface) is considerably below the kinetic energy with which the electron is ejected; yet even under these conditions the electrons are being emitted with the same kinetic energy.

The rather remarkable law which relates the kinetic energy w with which the electrons are emitted to the frequency ν of the incident radiation is

$$w = h\nu - w_0, \tag{3}$$

where h is Planck's constant and w_0 is a constant characteristic of the metal in question. For values of ν which would make the right-hand side negative (that is for $\nu < \nu_0$, where $\nu_0 = w_0/h$), there is no emission. Thus the photoelectric effect exhibits the curious property of setting in suddenly at a certain frequency ν_0 which is called the *photoelectric threshold* and is, like w_0, a characteristic of each metal. This law also obstinately frustrates every attempt at classical justification, although the presence within it of the same constant h which occurs in the theory of the black body gives an indication of its relation to some profound law of nature.

Einstein, in a celebrated paper,[5] showed that the strange behavior of the photoelectric effect is consistent with a corpuscular model of radiation. He considered this model solely for its heuristic value, however, without emphasizing its concrete significance. Indeed, if it is assumed that the energy of monochromatic radiation is not uniformly distributed over the entire wave front but instead travels localized in corpuscles (called light *quanta* by Einstein, also called *photons*[6] today) each of which contains a quantity of energy $h\nu$ and travels with the velocity of light, the difficulties of the photoelectric phenomenon disappear and we

---

[5] "Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt." *Ann. d. Physik* **17**, 132 (1905).

[6] This designation is not to imply that the hypothesis applies only to light, strictly speaking; it includes every type of electromagnetic radiation.

arrive immediately at formula (3). As a matter of fact, the paradox concerning photoelectric emission with weak light is explained right away when we consider that when the light is weak, there will be few light quanta but each will always contain the energy $h\nu$. Hence there will be few atoms receiving a quantum, but each of those few atoms will always receive an energy $h\nu$, independently of the light intensity. Formula (3) is then immediately justified if we interpret w_0 as the energy the electron has to expend to break away from the metal. Indeed, the electron, which has received the energy $h\nu$ from the light quantum, will, upon expending an amount w_0 of this energy to escape from the metal, emerge with a residual energy $h\nu - w_0$ in the form of kinetic energy. If the energy of the light quantum is below w_0 (that is, if $\nu < \nu_0$), the electron will not be able to escape, as has been actually found.

It is to be noted that the hypothesis of "light quanta" is not exactly a return to the old corpuscular (or ballistic) theories of light, since the energy of light quanta does not necessarily have to be thought of as kinetic energy of a particle in motion. Attempts have been made to interpret the energy of light quanta as electromagnetic energy localized in a small region of space, but it is not possible to reconcile such a view with the classical equations of Maxwell.

We may observe, however, that the hypothesis of "light quanta" is distinct from the hypothesis of the quanta of Planck: the former deals with the manner in which energy travels through space; the latter refers to the manner in which energy is emitted and absorbed by matter. Nevertheless, it is evident that these hypotheses strengthen each other, and that both groups of phenomena to which they refer make it clear that a profound physical relation must exist between the frequency ν of a radiation and the quantity $h\nu$ of energy.

As will be said later on, the hypothesis of light quanta has subsequently been confirmed by other experimental facts, and in present theory the concept of *light quantum* or *photon* plays as important a part as that of the electron and the proton.

It is known that the electromagnetic theory of light proves that with a radiant energy W there must be associated an *electromagnetic momentum* W/c, which manifests itself, when absorbed by matter, by the phenomenon of *light pressure* (cf. bibliography 32a, §142 and

§143). Hence it is natural that every theory of light, in order to be in accord with the facts, must associate a momentum W/c with radiant energy W. Thus, in addition to an energy $h\nu$, we must attribute to photons, independently of any concrete corpuscular concept, a momentum **p** of magnitude

$$p = \frac{h\nu}{c} = \frac{h}{\lambda}$$

in the direction of propagation of the light.[7] A further experimental proof of the exactness of this hypothesis is furnished by the Compton effect and by certain experiments related to it, which we shall now discuss briefly.

9. The Compton effect. When a beam of X rays traverses matter, part of it will be scattered in all directions in the manner of light which traverses a turbid medium. Analysis of this scattered[8] radiation with a crystal spectrograph demonstrates that it does not possess a spectral constitution identical to the primary radiation. If, for instance, the latter is monochromatic of frequency ν, the spectrum of the scattered radiation will exhibit, in addition to the line of frequency ν, a line of slightly lower frequency ν'. This line has considerable intensity, even superior to the intensity of the "undisplaced" line, if the scattering substance is of low atomic number (such as carbon or aluminum) and if the incident rays are "hard" (high ν), whereas it is rather weak or is missing altogether if the substance has high atomic number (such as silver) and if the incident rays are soft. This phenomenon of scattering with a slight increase of wavelength was discovered in 1923 by the American physicist A. H. Compton and carries his name.

The change in wavelength is different for radiation scattered in the different directions: it is zero in the direction of the primary rays, increases with the angle θ between the direction of observation and

---

[7] We remember from relativistic mechanics (cf. §§184 *ff* of No. 32a of the Bibliography) that a point particle having a *rest mass* m_0 has, at velocity v, a mass $m = m_0/\sqrt{1 - (v^2/c^2)}$, a momentum $p = mv$, and a kinetic energy $w = (m - m_0)c^2$. These formulas can also be applied to the photon, provided that we set $v = c$, $m = h\nu/c^2$, $m_0 = 0$. Thus we can think of the photon as a limiting case of a point particle, having a rest mass zero and a velocity equal to c.

[8] With the actually scattered radiation there is mixed in the *secondary radiation* whose spectrum is the one characteristic of the scattering substance; this is easily distinguished from the scattered one, and we shall not be concerned with it here.

that of the primary rays, and reaches its maximum value (which amounts to only 0.048 Å) for $\theta = 180°$, that is, when the scattered radiation is observed straight back. This variation is well represented by the formula

$$\lambda' = \lambda + 0.024(1 - \cos\theta), \tag{4}$$

where λ and λ' must be expressed in angstrom units. As can be seen from this formula, the change in wavelength $\lambda' - \lambda$ is independent of λ, so that the *relative* effect is appreciable for small values of λ, or for "hard" rays, whereas it becomes negligible when λ is large compared with 0.048 Å.

The classical theories give no explanation of the Compton effect, since according to these theories the scattering is due to the fact that the electrons contained in the scattering substance execute forced oscillations under the influence of the alternating electric field of the incident radiation, and hence in turn become centers of emission for waves. However, these oscillations must take place with the frequency of the incident radiation, and hence the scattered radiation must have precisely the same frequency.

10. Quantum theory of the Compton effect. According to the theory of quanta, the scattering process must instead be thought of as the effect of collisions between photons of the incident radiation and the electrons of the scattering substance. After collision, the photons are deviated from their original direction just like the alpha particles in the Rutherford experiment which we have already discussed. This collision, no matter what its mechanism, is governed by the two fundamental laws of conservation of energy and momentum, as if we were dealing with a collision between perfectly elastic bodies. Hence it is clear that the electron acquires a certain velocity in the impact and thus removes a certain amount of energy from the incident photon, which is therefore scattered with less energy. And since the energy of the photons is related to their frequency by the relation $E = h\nu$, the scattered photon must have a frequency lower than that of the incident photon. In this way the Compton effect is explained qualitatively.

We may now set down this reasoning quantitatively. Let us call v the velocity acquired by the electron that has been struck; then its kinetic energy will be, according to relativistic mechanics,

$mc^2(1/\sqrt{1-(v^2/c^2)}-1)$, where m is the *rest mass* of the electron);[9] whereas the energy of the incident quantum is $h\nu$ and the energy of the scattered quantum is $h\nu'$. The principle of conservation of energy yields

$$h\nu = h\nu' + mc^2\left(\frac{1}{\sqrt{1-(v^2/c^2)}} - 1\right) \tag{5}$$

Let us take three vectors (see Fig. 3) to represent, respectively, the momentum of the incident photon (vector AO, of length $h\nu/c$), the momentum of the scattered photon (vector OB, of length $h\nu'/c$), and the momentum acquired by the electron (vector OD, of length $mv/\sqrt{1-(v^2/c^2)}$). We observe that the first vector must be the resultant of the other two. From the triangle OBC, using the cosine law, we obtain immediately

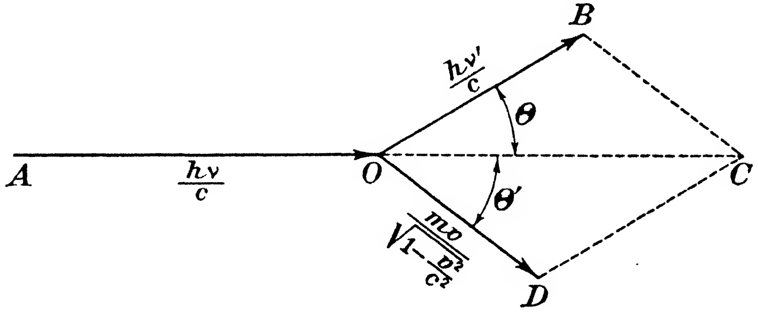

Fig. 3

$$\frac{m^2v^2}{1-(v^2/c^2)} = \frac{h^2\nu^2}{c^2} + \frac{h^2\nu'^2}{c^2} - 2\frac{h\nu}{c}\frac{h\nu'}{c}\cos\theta. \tag{6}$$

Eliminating v between equations (5) and (6) gives

$$\nu - \nu' = \frac{h\nu\nu'}{mc^2}(1-\cos\theta). \tag{7}$$

By expressing the frequencies in terms of the wavelengths, that is, setting $\nu = c/\lambda$, $\nu' = c/\lambda'$, we obtain, after simplifying,

$$\lambda' = \lambda + \frac{h}{mc}(1-\cos\theta). \tag{8}$$

---

[9] If we had used the classical expressions for kinetic energy and momentum, we should have arrived at results which are practically indistinguishable from those reached by relativistic mechanics, but the formulas would be somewhat more complicated.

If we substitute the numerical values for h, m, and c, we obtain

$$\frac{h}{mc} = 2.43 \times 10^{-10}\,\text{cm} = 0.0243\ \text{Å},$$

so that equation (8) coincides exactly with (4), which describes, as has been said, all experimental results.

In this theory it has been assumed that the scattering electrons are free. In reality they are more or less bound to the nucleus of the atom to which they belong, and hence the entire atom (and possibly even an entire molecule) participates to a certain extent in the collision. In order to see the effect of this circumstance, we shall consider the extreme case where the electron is rigidly bound to its atom. It will then behave like a body, not of mass m but of mass M equal to the mass of the atom, that is, several thousand times larger. The preceding reasoning is still valid, provided that we substitute M for m. Hence in the final formula (8), the constant h/mc, which determines the entire effect, will become several thousand times smaller, which means that the Compton effect will become negligible. This consideration justifies the presence of the *undisplaced line*, which is due to the electrons more strongly bound to the nucleus, whereas the displaced line is due to those peripheral electrons which can be considered to be practically free. For a given frequency ν in elements of high atomic number, that is, with high nuclear charge, it is natural that the effect of the bound electrons (which are more numerous) should prevail, whereas in the light elements almost all the electrons behave as if they were free. It is also clear that, in a given substance, the electrons can be considered the freer, the higher the energy of the protons that impinge upon them, or else the higher the frequency ν. As has been said, all this is in perfect agreement with experience.

11. The experiment of Bothe and Geiger. The preceding theory of the Compton effect (which is due to Compton and Debye) received valid support in an experiment performed by Bothe and Geiger in 1925.[10] It purported to verify whether, simultaneously with the scattering of the photon, an electron is actually projected (*knock-on electron*), as the theory predicts. In order to detect both the scattered photons and the knock-on electrons, these investigators

---

[10] W. Bothe and H. Geiger, *Zeits. f. Physik* **32**, 639 (1925).

used a *point counter*, an instrument that had already been employed by Rutherford and Geiger to count the alpha particles of radium. It consists of a special ionization chamber capable of giving an electrical signal whenever a single electron or photon or any ionizing particle enters it. In the experiment of Bothe and Geiger, two point counters were placed on either side of a narrow beam of X rays traversing hydrogen. One of the two counters was covered with a thin platinum foil able to stop the electrons but not the X rays; hence this counter registered only the photons scattered by the hydrogen. The other counter registered the knock-on electrons, and in addition the photons scattered to its side of the beam, as well as the electrons occasionally produced by the photoelectric effect. Therefore a complete coincidence between the signals from the two counters could not be expected, and in fact the second counter gave quite a few more signals than the first. However, the experimenters observed the remarkable result that for each signal of the first counter there was a *simultaneous* one in the second counter—proof that for each photon scattered to one side there is always a knock-on electron propelled to the other side, in perfect coincidence.

12. The experiment of Compton and Simon. An even more significant experiment for the confirmation of the preceding theory of the Compton effect was performed by Compton and Simon in 1925.[11] They used the *cloud-chamber method* of Wilson, which permits the trajectories of the electrons traversing a gas to be made visible in the form of fine droplet tracks. In this manner these investigators succeeded not only in revealing the existence of the knock-on electron photographically but also in verifying that the direction in which it is projected is the one predicted by the laws of elastic collision.

We shall now see what relations between the direction OB of the scattered photon and the direction OD of the knock-on electron are given by these laws, that is, between the angles θ and θ' (Fig. 3). Instead of writing down relations (6) between only the magnitudes of the three vectors in question, we shall use the principle of conservation of momentum in full by expressing the fact that the projection of the momentum in the direction AO and its projection in a direction perpendicular to AO must be conserved. We then write, instead of (8), the two equations

[11] A. H. Compton and A. W. Simon, *Phys. Rev.* **26**, 289 (1925).

$$\left.\begin{aligned} \frac{h\nu}{c} &= \frac{h\nu'}{c}\cos\theta + \frac{mv}{\sqrt{1-(v^2/c^2)}}\cos\theta', \\ 0 &= \frac{h\nu'}{c}\sin\theta - \frac{mv}{\sqrt{1-(v^2/c^2)}}\sin\theta', \end{aligned}\right\}$$

whence (leaving only the last term on the right-hand side and dividing the second by the first equation)

$$\frac{\nu'\sin\theta}{\nu-\nu'\cos\theta} = \tan\theta'.$$

If we then substitute for ν' the expression obtained from (7), we have, after a simple transformation,

$$\tan\theta' = \frac{1}{1+(h\nu/mc^2)}\cot\frac{\theta}{2}. \tag{9}$$

This is the desired relation. From it we find, in particular, that when we vary θ from 0 to 180°, θ' varies from 90° to 0°. Hence the electrons are always scattered forward.

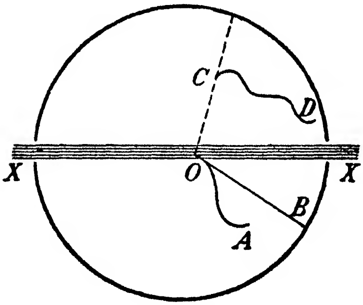

Fig. 4

In order to verify this formula experimentally, Compton and Simon took a large number of stereoscopic photographs with the cloud chamber containing a gas traversed by a narrow beam of X rays. Many of these photographs caught a scattering act and appear as shown schematically in Fig. 4, where XX represents the beam of primary X rays. From a point O of the beam there starts the track OA of a knock-on electron. The scattered photon cannot leave a track along its trajectory; it becomes visible when it encounters an atom C which absorbs it, since that atom then emits an electron by the photoelectric effect, and the track CD of the electron is visible in the photograph. By connecting its origin C with O, we obtain the direction in which the photon has been scattered (dashed line). The tangent OB to the track OA, drawn near its origin, gives the initial direction of the knock-on electron. By means of the two stereoscopic photographs we can thus determine the values (in space) of the angles θ and θ', and it is found that they satisfy equation (9) within the limits of experimental error.

More recently, Crane, Gaertner, and Turin performed analogous experiments with gamma rays (instead of X rays), measuring directly (by deflection in a magnetic field) the velocity of the knock-on electron, as well as the angles θ and θ'. These experiments are also in accord with theory.

13. Impossibility of a literally corpuscular theory of light. The hypothesis that light is of a corpuscular nature (that it is composed of particles in the intuitive sense of the word) seems *almost* to be dictated by the phenomena of which we have spoken in the preceding sections. Nevertheless, it encounters serious difficulties in another and no less vast category of phenomena: those which have constituted, from the time of Huygens on, the experimental basis of the wave theories of light. The most important of these phenomena are interference and diffraction.

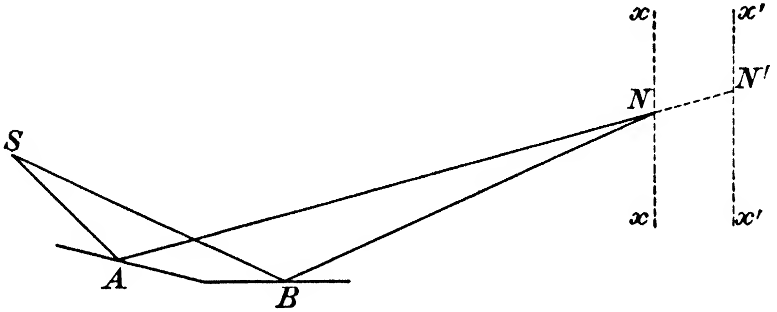

Fig. 5

In order to realize the nature of these difficulties, let us consider a particularly simple case of interference, that of the Fresnel mirrors (Fig. 5). Let S be the source, A and B the mirrors, and N a point such that the optical paths SAN and SBN differ by an odd number of half wavelengths. Let us suppose at first that mirror B is covered by a screen. Then there will be light at N, or else (using corpuscular terminology) particles which have followed the line SAN will arrive at N. If we uncover mirror B, a dark fringe is produced at N, whereas at other points the illumination is reinforced (bright fringes). We might suppose that the particles that came along the line SAN are being influenced, from a distance, by the covering up of mirror B and are being deviated. However, this hypothesis, besides being strange in itself, does not explain why, upon observing the fringes in another plane $x'x'$ instead of plane xx, we may find a bright fringe at N', that is, on the extension of ray AN. We must, then, suppose that the particles coming from A have passed through N but have been rendered ineffective in some manner at that point by the simultaneous presence of particles coming from B. But this hypothesis, besides leaving unexplained the fact that in other points of the screen there is reinforcement of illumination, is contradicted by the experimental fact that the phenomena of interference are still taking place when the light is so weak that the apparatus is traversed by one particle at a time. Thus we are forced to repudiate the assumption that the particles may interact in such a manner as to render each other ineffective in certain points and to reinforce each other at other points.

A proof that interference occurs even with only one quantum at a time can be found in the fact that astronomical telescopes give clear images even of the very weakest stars, and it is known that images in optical systems are formed by a process of interference. Stars of 10th magnitude are perfectly visible with a telescope of 10 cm aperture; their radiation has an intensity of about 3.5×10^{-10} erg/cm²/sec.

Assuming light of an average wavelength $\lambda = 5500$ Å and calculating, the energy corresponding to each quantum by using the relation $E = h\nu$ we find 36×10^{-13} erg; hence the intensity of a 10th-magnitude star is approximately 100 quanta per cm² per second. There are thus 8000 quanta entering the telescope every second. If we consider that each quantum traverses the telescope in a time of the order of 10^{-8} sec, we see that the individual quanta traverse the telescope separately, with large intervals between them. The image formation under these conditions would imply that each single quantum uses the entire aperture of the objective in turn, and would lead us to attribute to the quanta transverse dimensions of many centimeters or several meters, in contrast with the experiments on the photoelectric effect, which led us to assign to them atomic dimensions.

The interference of single quanta has further been proved by direct experiments of G. J. Taylor[12] and of Dempster and Batho[13]; the latter have obtained, by means of an echelon interferometer, interference fringes even when the light was so weak that certainly each quantum entered the apparatus after the preceding one had left it.

In conclusion, every theory of light which is corpuscular in the literal sense of the word, that is, which attributes *definite trajectories*

---

[12] *Proc. Camb. Phil. Soc.* **15**, 114 (1909).

[13] *Phys. Rev.* **30**, 664 (1927).

to light quanta as when dealing with material particles, has turned out to be absolutely irreconcilable with experimental facts, just as the wave theory of Maxwell has proved to be in contrast with the photoelectric effect and with many other experimental facts. We shall see at a later stage (§34) how a profound revision of the principles of the wave theory and the corpuscular theory has succeeded in eliminating this contrast by establishing a unified basis and leaving to each of the two old theories that portion of validity which is in accord with experimental facts.

CHAPTER 3

Energy Levels

14. The spectrum of hydrogen and of hydrogenlike systems. Spectroscopy, in addition to furnishing the largest amount of information about atoms, is characterized by very high precision in yielding wavelength measurements, which can often be performed with a relative error below 1 part per million.

Of all the spectra, the first to have been interpreted was that of atomic hydrogen. This spectrum is the simplest, in accordance with the fact that the hydrogen atom has the simplest structure.

If we take a spectroscope and observe the light emitted by a Plücker tube containing hydrogen, we perceive a large number of lines that can easily be classified into two categories. There is a background made up of numerous and closely spaced lines of low intensity. In contrast are a small number of easily visible lines, whose intensity increases considerably faster than that of the background when the current through the tube is raised. It has been ascertained that the numerous weak lines (which constitute the so-called *many-line spectrum*) are emitted by the H_2 *molecules*, whereas the more intense ones (which form the so-called *Balmer spectrum*) are due to the H *atoms*, which are produced in the tube by the effect of dissociation caused by the passage of current. The many-line spectrum, being emitted by the molecules, has a complicated structure and has only recently been interpreted theoretically; we shall concern ourselves here only with the Balmer spectrum, referring to it exclusively when speaking of the hydrogen spectrum.

This spectrum is made up of lines partly in the infrared region, partly in the visible region, and partly in the ultraviolet region. It is represented schematically in Fig. 6, which assumes a dispersion proportional to the frequency. As is evident from the figure, these lines are clustered into three groups, called *Paschen series* (infrared), *Balmer series* (visible) and *Lyman series* (ultraviolet), respectively.

The first step toward the theoretical interpretation of spectra was made in 1885 by the Swiss scientist Balmer, who observed that all the lines of the series of hydrogen which carries his name (and which was the only one known then) have frequencies[1] that can be represented by the formula

$$\tilde{\nu} = R\left(\frac{1}{4} - \frac{1}{n^2}\right), \tag{10}$$

in which R is a constant, now called the *Rydberg constant,* whose numerical value is $R = 109{,}678\ \text{cm}^{-1}$, and n is an integer which can take on all values from 3 on. As can readily be seen, when the value of n increases, the values of $\tilde{\nu}$ given by formula (10) tend toward the limit $R/4$; in fact, the lines of the Balmer series condense toward a limit, so that beyond a certain point it is no longer possible to distinguish between them. Nevertheless, more than 30 of them have been resolved, for all of which the frequency very exactly satisfies Balmer's formula.

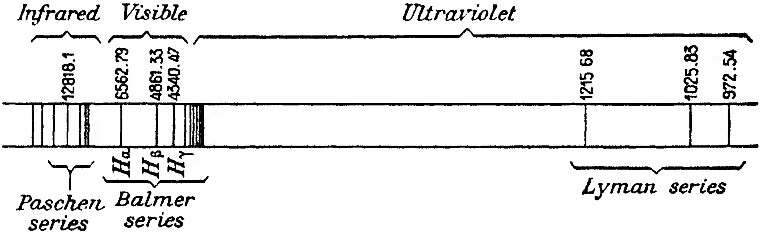

Fig. 6

---

[1] In spectroscopy it is an established practice to call "frequency" of a radiation not only the true frequency ν (number of vibrations per second, related to the wavelength by $\nu = c/\lambda$) but also a quantity proportional to it:

$$\tilde{\nu} = \frac{\nu}{c} = \frac{1}{\lambda}.$$

This is also more properly called "wave number," since it represents the number of wavelengths in one centimeter. The use of "wave number" is generally preferred to the use of frequency itself, since λ can be measured directly with great precision, sometimes better than 1:1,000,000; it is sufficient to take its reciprocal to get $\tilde{\nu}$, whereas ν has to be obtained from λ by the formula $\nu = c/\lambda$. Since c is known with a precision little above 1:100,000 ($c = 2.99776 \times 10^{10}$ cm/sec), ν turns out to be known with somewhat lower precision than $\tilde{\nu}$.

To avoid misunderstandings, it will suffice to add to the designation of "frequency" the unit of measurement, which is sec^{-1} for the true frequency ν, and is cm^{-1} for wave numbers $\tilde{\nu}$. Generally, the wave number $\tilde{\nu}$ is also indicated by the simple letter ν.

In the following discussion, it will be recognized that Balmer's formula is merely a particular case of a more general formula representing all the lines of the spectrum of atomic hydrogen. This formula is

$$\tilde{\nu} = R\left(\frac{1}{n'^2} - \frac{1}{n^2}\right), \tag{11}$$

where R is the same constant as above, and n' and n are two integers. Taking $n' = 1$ and $n = 2, 3, 4, \ldots$, we obtain the frequencies of the *Lyman series:*

$$\tilde{\nu} = R\left(1 - \frac{1}{n^2}\right).$$

Taking $n' = 2$ and $n = 3, 4, 5, \ldots$, we again obtain (10), which represents the *Balmer series.* Finally, taking $n' = 3$ and $n = 4, 5, 6, \ldots$, we obtain the frequencies of the *Paschen series:*

$$\tilde{\nu} = R\left(\frac{1}{9} - \frac{1}{n^2}\right).$$

If we take $n' = 4$, we obtain another infrared series, called the *Brackett series,* of which only a few lines are known.

The simplicity of formula (11), the remarkable number of lines which it represents, and the extreme precision with which it fits experimental results practically rule out the possibility of accidental coincidence. Hence, attempts were made to interpret the formula with an appropriate atomic model, but for a long time these encountered apparently insuperable difficulties.

Later on, the discovery was made that spectra strictly analogous to the hydrogen spectrum are emitted by atoms of the lighter elements, when they are ionized in such a way that they have lost all their electrons but one—that is, by the ions He^+, Li^{++}, Be^{+++}, B^{++++}. These ions, being made up of a nucleus and a single electron, have a structure analogous to the hydrogen atom, from which they differ only in the mass of the nucleus and in its charge, which, instead of being e as in the hydrogen atom, is Ze for the element of atomic number Z. These systems are collectively termed *hydrogen-like systems.* It can be foreseen that to this structural analogy there must correspond an analogy of spectral properties; in fact, there is reason to suppose that the spectrum of an atom of atomic number Z, ionized $(Z - 1)$ times, will be represented by a formula analogous to (11):

$$\tilde{\nu} = Z^2R\left(\frac{1}{n'^2} - \frac{1}{n^2}\right) \tag{11'}$$

and hence will differ from the hydrogen spectrum only because the frequencies of all lines are multiplied by Z^2 (except for a slight correction due to the movement of the nucleus, of which we shall speak in §58 of Part II). For example, the He^+ ion $(Z = 2)$ emits a series called the *Pickering series,* given by

$$\tilde{\nu} = 4R\left(\frac{1}{4^2} - \frac{1}{n^2}\right), \tag{12}$$

whose frequencies are four times those of the infrared (*Brackett*) series of hydrogen and fall in the region of visible light rather than in the infrared. Several lines of this series have been observed;[2] also some lines of Li^{++}, Be^{+++} and B^{++++}, all satisfying formula (11′).

Finally we shall point out the fact that the lines of hydrogen and those of hydrogenlike ions, upon observation with instruments of high resolving power, each prove to be composed of a group of several very closely spaced lines which constitute the so-called *fine structure* of the lines. In hydrogen the difference in wavelength of the various components is, at most, of the order of a few tenths of an angstrom unit and hence is detectable only by methods of high resolution. In ionized helium, however, it reaches several angstroms, and therefore the fine structure of these lines can be detected even with a grating. We shall turn to this phenomenon later on in order to give its theoretical explanation.

Many lines of other spectra further show a so-called *hyperfine structure,* which requires means of very high resolution and has an entirely different origin from the preceding fine structure.

---

[2] The Pickering series, first observed only in the spectrum of a star, was attributed to hydrogen, because its lines of even position coincide with the lines of the Balmer series, as can be seen by observing that (12) can also be written

$$\tilde{\nu} = R\left(\frac{1}{4} - \frac{1}{(n/2)^2}\right).$$

Subsequently, since the Bohr theory predicted that He^+ must emit such a series, these lines were examined again and were found in the spectrum emitted by a tube containing helium which was absolutely devoid of hydrogen.

15. Spectral series. Spectroscopic terms. Guided by the discoveries of the hydrogen series, the spectroscopists started to look for analogous regularities in the spectra of other elements, and in several other cases they succeeded in finding that the lines are grouped into *series*. The lines of the same series are characterized by a general appearance similar in structure, intensity, and sharpness, and by an analogous behavior with respect to various physical influences (such as magnetic fields) that may act upon the source. For many series they also found empirical formulas analogous to the Balmer formula.

In all these formulas—just as in Balmer's—the frequency of a line occurs as the difference between two *terms*, of which the first one remains constant, whereas the second assumes different values (corresponding to the successive integers) for the different lines of a given series. The terms may quite often be represented by the formula (given by Rydberg)

$$\tau_n = \frac{R}{(n - \alpha)^2}, \tag{13}$$

where R is the Rydberg constant and α another constant. By holding α fixed and assigning to n all integral values from a certain value on up, we obtain a *sequence* of terms. In particular, this formula contains the *Balmer terms* (for which $\alpha = 0$):

$$\tau_n = \frac{R}{n^2}.$$

A given element usually presents different sequences of terms, each one being characterized by a given value of α; and the frequencies of a series of lines are obtained in general by combining a fixed term of a sequence with all the terms of another sequence. Thus the formula that yields the frequencies of a series is of the type

$$\tilde{\nu} = R\left[\frac{1}{(n' - \alpha')^2} - \frac{1}{(n - \alpha)^2}\right],$$

where n' is fixed and n assumes all integral values above a certain value.

However, for many elements the terms τ_n have a more complex form than that of Rydberg. A second type, for instance, is represented by the so-called Ritz formula

$$\tau_n = \frac{R}{(n - \alpha + K\tau_n)^2},$$

where K is another constant characteristic of each sequence.

Finally, in many cases, though it may not be possible to represent the terms τ_n by simple formulas, the frequencies of the various spectral lines may nevertheless be written in the form

$$\tilde{\nu} = \tau_{n'} - \tau_n,$$

thus causing many frequencies of the same substance to depend on a rather limited number of *terms*. Hence we can say that the first step in the interpretation of the spectrum of a substance consists in searching, within the list of frequencies of its spectral lines, for the more limited list of "terms" which, when combined, give rise to the frequencies. As we shall see in the following section, the terms of a spectrum have an important physical significance not possessed by the frequencies of the single lines.

Since, by taking the difference of a relatively small number of terms, we can arrive at a considerably larger number of frequencies, it is evident that large numbers of lines of constant frequency difference must exist in a spectrum. This fact, which was discovered empirically by Ritz, carries the name of *combination principle*. By suggesting the search for terms, it furnishes the key for the interpretation of complex spectra.

16. Note on the Bohr theory of the hydrogen atom. The fact that the frequencies of spectral lines can generally be represented by the difference of two terms received, in 1913, an interpretation by Bohr[3] which has since been confirmed by other experimental facts and has proved to be very fertile; it still constitutes the basis of spectroscopy. Bohr proposed this theory to interpret the Balmer series and related ones (to which we shall refer in our explanation of the theory), but his fundamental concept of *energy levels* can be extended to all spectra.

According to the Rutherford model, the hydrogen atom consists of a nucleus and of an electron which describes a Keplerian ellipse about it, or, in particular cases, a circle. For simplicity, we shall refer here to the circular case, as does Bohr; and since the theory applies to all hydrogenlike systems, we shall immediately consider a hydro-

---

[3] *Phil. Mag.* **26**, 1 (1913).

genlike system of atomic number Z. It is known from mechanics that the radius r of the orbit is determined by the initial conditions of the motion and can have any value whatever, so long as the velocity v is such that the centrifugal force equals the electrostatic attraction of the nucleus, or $mv^2/r = Z\,e^2/r^2$, so that v is related to r by

$$v = \sqrt{\frac{Z\,e^2}{mr}}. \tag{14}$$

The energy of the system (kinetic energy plus potential energy) can then be expressed as a function of r alone, and we get

$$E = \frac{1}{2}mv^2 - \frac{Z\,e^2}{r} = \frac{1}{2}\frac{Z\,e^2}{r} - \frac{Z\,e^2}{r} = -\frac{1}{2}\frac{Z\,e^2}{r}. \tag{14'}$$

(The negative value of E signifies that it is necessary to do work in order to break up the atom by removing the electron to an infinite distance and bringing it to rest—that is, in order to ionize the atom without imparting any velocity to the electron.)

Now Bohr assumed that in atomic mechanics there exists a supplementary condition, according to which motion is not possible on all the circles permitted by ordinary mechanics but only on some of them, called "stable orbits" or "quantum orbits," whose radii constitute an infinite but "discrete" (that is, not continuous) sequence $r_1, r_2, \ldots$. This supplementary condition was formulated by Bohr as follows. The *angular momentum* (or moment of momentum) $M = mrv$ has to be an integral multiple of $h/2\pi$:

$$M = n\frac{h}{2\pi} \quad (n = 1, 2, 3, \ldots). \tag{15}$$

Now, because of (14),

$$M = \sqrt{mrZ\,e^2},$$

so that, upon substituting into the preceding expression and solving for r, we find that the radius of the nth orbit, which we shall indicate by a_n, is

$$a_n = n^2\frac{h^2}{4\pi^2\,e^2mZ}. \tag{16}$$

Figure 7 represents the first three quantum orbits. The innermost orbit has (in the case of hydrogen, where $Z = 1$) a radius

$$a_1 = \frac{h^2}{4\pi^2\,e^2m} = 0.527 \times 10^{-8}\text{ cm}.$$

This value turns out to be of the order of magnitude required for atomic dimensions by the kinetic theory of gases. The successive orbits have respective radii 4 times, 9 times, and so on, times the radius of the first orbit. These orbits represent, so to speak, imaginary tracks upon which the electron is constrained to move, according to the Bohr theory.

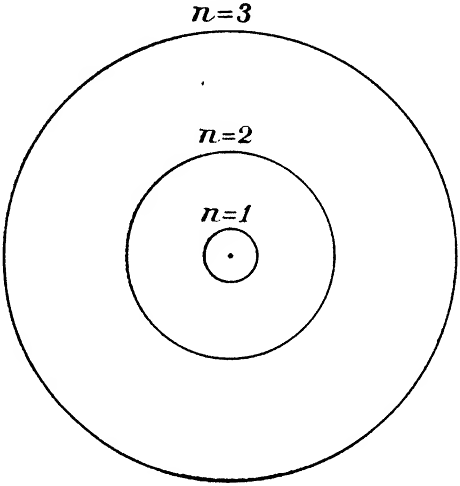

Fig. 7

To each of these privileged orbits there corresponds an energy

$$E_n = -\frac{1}{2}\frac{Z\,e^2}{a_n}.$$

Thus the energy of the atom can assume only certain *discrete* values E_n, called *energy levels*, each of which corresponds to a different "state" of the atom (*quantum state*), characterized, according to Bohr, by the motion of the electron on one or another of the quantum orbits. From the expression (16) for a_n, we find that these energy levels are given by

$$E_n = -Z^2\frac{2\pi^2\,e^4m}{h^2}\frac{1}{n^2}. \tag{17}$$

Bohr assumed further that the electron can pass, with a sudden jump (*quantum jump*), from one orbit (n) to another (n'). Bohr makes no assumption on the nature of that passage, except that it is to obey the principle of conservation of energy. Therefore the atom must *absorb* or *emit* the energy corresponding to the difference between E_n and $E_{n'}$ (*absorb* if the transition occurs from an internal to an external orbit, *emit* in the opposite case). This energy is ordinarily absorbed or emitted in the form of radiation. In the Bohr theory, the mechanism of emission and absorption is thus not at all the one of ordinary electromagnetic theory (based on the fact, governed by Maxwell's equations, that an electron radiates if it does not move in uniform rectilinear motion). It is even explicitly being denied that this theory holds in the atomic domain, since otherwise emission would take place continuously during the motion of an electron on a quantum orbit. Hence the energy of the atom would continue to diminish gradually, a condition irreconcilable with the existence of discrete energy levels. Instead the emission occurs, so to speak, in spurts, every time an electron passes from one quantum orbit (n) to a lower-lying one (n'), and no assumption is made concerning the mechanism of this emission. The only assumption concerns the frequency of the emitted radiation. This hypothesis is suggested by the fact, already discovered by Einstein (cf. §8), that in various phenomena (the photoelectric effect, for example) the radiation of frequency ν is assembled into "quanta," each containing an energy $h\nu$. By identifying each of these quanta with a burst of radiation emitted in a quantum jump, we are led to assume that the *quantity* $E_n - E_{n'}$ of emitted energy and its *frequency* ν (in $\sec^{-1}$) are connected by the relation

$$\nu = \frac{E_n - E_{n'}}{h}, \tag{18}$$

which is still the fundamental relation of spectroscopy. Analogous relations hold for absorption (except for the interchange of E_n and $E_{n'}$).

From (18) we immediately see why the frequencies emitted by the atom occur as differences of "spectral terms," and at the same time we recognize the physical significance of the latter: each of the terms corresponds to a different energy level of the atom, and hence to a different quantum state. Thus if we divide (18) by c in order to go from ν to $\tilde{\nu}$ and if we set

$$\tau_n = -\frac{E_n}{hc}, \tag{19}$$

(note that in this way τ_n is positive), we obtain

$$\tilde{\nu} = \tau_{n'} - \tau_n = Z^2 \frac{2\pi^2 e^4 m}{ch^3} \left(\frac{1}{n'^2} - \frac{1}{n^2} \right), \tag{20}$$

which is of the same form as formula (11) for the Balmer and analogous series. Even more remarkably, when we insert for e, m, and h their numerical values obtained from other physical phenomena,[4] the coefficient $2\pi^2 e^4 m/ch^3$ of (20) becomes numerically equal (within the limits of experimental error) to the Rydberg constant R; in fact, we find that $R = 109{,}710\ \text{cm}^{-1}$ by this method, whereas experiment yields $R = 109{,}677.58\ \text{cm}^{-1}$. The agreement becomes even better when we take the correction due to the motion of the nucleus into account (see Part II, §58) which brings the theoretical value (for hydrogen) to 109,650 cm.$^{-1}$, with a difference from the experimental value of about 0.03 percent. The Bohr theory thus accounts quite well for all the more than 50 lines of atomic hydrogen and hydrogenlike ions, yielding for the Rydberg constant the expression

$$R = \frac{2\pi^2 e^4 m}{ch^3}. \tag{21}$$

Subsequently the Bohr theory was amplified and perfected, especially by Sommerfeld. By considering elliptical as well as circular orbits and by introducing various important refinements, Sommerfeld was able to account not only for the minutest details of the hydrogenlike spectra but also to open the path for the study of more complex atoms. Up until the rise of quantum mechanics in the late 1920's, this study developed very successfully on the basis of Sommerfeld's methods. He postulated, in the place of Bohr's condition (15) more general conditions able to characterize the quantum orbits even in cases which are more complex than that of Bohr (*Sommerfeld conditions*).[5] These conditions, which (just like

---

[4] The calculations are made by taking for e, m, and h the best values available in 1949 among those obtained by methods which do not make use of (21) [see R. T. Birge, *Rev. Mod. Phys.* **13**, 233 (1941)]: $e = 4.8025 \times 10^{-10}$ e.s.u. (from the faraday and Avogadro's number), $h = 6.624 \times 10^{-27}$ erg-sec (from the fine-structure constant), $m = 9.1066 \times 10^{-28}$ g (from e/m), all with probable errors of the order of 0.03 per cent.

[5] These were proposed independently and almost simultaneously by W. Wilson [*Phil. Mag.* **29**, 795 (1915)], Ishiwara [*Tokyo Math. Phys. Proc.* **8**, 106 (1915)], and A. Sommerfeld [*Ann. d. Physik* **51**, 1 (1916)].

Bohr's, which is a particular case thereof) originally appeared to be rather strange a priori postulates, justifiable only a posteriori by the (not always complete) correctness of their consequences, can today be deduced as results of first approximation of the Schrödinger theory. We shall postpone the presentation of the Bohr-Sommerfeld theory from that standpoint until Part II, proceeding now to examine those fundamental ideas of this theory which retain all their value, even in the new atomic mechanics.

16a. Determination of the Rydberg constant by means of the correspondence principle. Although the quantum laws, which are obeyed by electrons in atoms, differ from the "classical" laws of ordinary rational mechanics and from Maxwell's electromagnetic theory, it is clear that the difference must be the less relevant, the larger the mechanical system with which we are dealing, until it must become entirely negligible for systems of ordinary size. This idea has been precisely stated by Bohr in a postulate called *correspondence principle,* which has served as a valuable guide to find the laws of atomic mechanics, and which we shall illustrate in §63 of Part I. We shall anticipate this principle by a treatment in limited form but sufficient to apply it to the problem of the hydrogen atom, discussed in the preceding section. We shall find that the correspondence principle can replace the postulate of the angular momentum expressed by (15) and that it enables us to obtain, by a fairly direct way, the expression (21) for the Rydberg constant.

First of all we recall that, according to classical electromagnetic theory, if the motion of an electric charge is periodic of frequency ν_0, the magnetic field generated by it will also be periodic, and hence each of its components can be considered to be a sum of functions, sinusoidal in time, of frequencies $\nu_0, 2\nu_0, 3\nu_0, \ldots$ (development in a Fourier series; see §9 of Part II). Therefore the radiation will be composed of spectral lines having these frequencies. The first is called the *fundamental frequency,* and its multiples are called *harmonics;* of course the latter may be missing entirely or in part.

We shall now calculate the frequencies which would be emitted, according to classical laws, by the electron of a hydrogen atom by virtue of its motion in a circle of radius r. The period, or the time required to complete the orbit, is $2\pi r/v$; hence the fundamental frequency (which is the reciprocal of the period) will be, taking into account expression (14) with $Z = 1$,

$$\nu_0 = \frac{v}{2\pi r} = \frac{e}{2\pi\sqrt{m}}\frac{1}{r^{3/2}}.$$

The frequencies which would be emitted according to the classical theory are thus given by[6]

$$\nu_{\rm cl} = s\frac{e}{2\pi\sqrt{m}}\frac{1}{r^{3/2}} \qquad (s = 1, 2, 3, \ldots). \qquad (\alpha)$$

Let us now take up the problem from the quantum point of view, and let us therefore introduce the following two postulates:

(a) There exist quantized orbits of radii $r = a_n (n = 1, 2, 3, \ldots)$, which we shall not characterize here. Their energies, by virtue of (14′), are given by

$$E_n = -\frac{1}{2}\frac{e^2}{a_n}. \qquad (\beta)$$

(b) The frequency emitted in the quantum jump from the nth to the n'th level is given by the Bohr formula (18).

For these frequencies to coincide with the experimental ones given by (11), we must have

$$E_n = -\frac{Rhc}{n^2};$$

and hence, because of (β),

$$a_n = \frac{e^2}{2Rhc}n^2. \qquad (\gamma)$$

This is the law which fixes the quantum orbits. The value of the constant R still has to be determined theoretically.

(c) With an increase of n and n' (and hence of the dimensions of the orbits), the frequencies given by the quantum theory must tend to become identical with the ones of the classical theory: this is the correspondence principle which will permit us to find the theoretical value of R.

Quantum theory yields, for the frequency emitted in the jump from the nth to the n'th orbit, the following expression (where we have set $n' = n - s$):

---

[6] In the case of circular motion, which we are here considering, the intensity of the harmonics turns out to be zero. This fact has no importance in the reasoning that follows, which can thus also serve as a model for the more general cases.

$$\nu_{\mathrm{qu}} = Rc\left[\frac{1}{(n-s)^2} - \frac{1}{n^2}\right] = \frac{Rc}{n^2}\left[\left(1 - \frac{s}{n}\right)^{-2} - 1\right].$$

If we develop the binomial in a power series and indicate by dots the terms with higher powers of s/n, the formula becomes

$$\nu_{\mathrm{qu}} = \frac{2Rcs}{n^3}\,(1 + \ldots).$$

On the other hand, the frequencies emitted according to the classical theory are given by (α), which, upon setting $r = a_n$ and using (γ), become

$$\nu_{\mathrm{cl}} = \sqrt{\frac{2}{m}}\,\frac{(Rhc)^{3/2}}{\pi\, e^2}\,\frac{s}{n^3}.$$

The ratio of these two expressions is

$$\frac{\nu_{\mathrm{qu}}}{\nu_{\mathrm{cl}}} = \frac{\pi\, e^2}{h^{3/2}}\sqrt{\frac{2m}{Rc}}\,(1 + \ldots).$$

If we let n tend toward infinity (s being held fixed) the equation tends to the limit (since the terms indicated by dots approach zero):

$$\lim_{n\to\infty} \frac{\nu_{\mathrm{qu}}}{\nu_{\mathrm{cl}}} = \frac{\pi\, e^2}{h^{3/2}}\sqrt{\frac{2m}{Rc}}.$$

Now the correspondence principle requires that this limit be equal to unity; thus we find that R must be given by

$$R = \frac{2\pi^2\, e^4 m}{ch^3}.$$

This expression for the Rydberg constant coincides with (21), but we have now found it without making use of condition (15) concerning the angular momentum; instead we have used the correspondence principle.

We can then easily deduce condition (15) by observing that from (14) (with $Z = 1$) we have $M = e\sqrt{mr}$. Substituting for r the value a_n given by (γ), we obtain

$$M = n\, e^2\sqrt{\frac{m}{2Rhc}}.$$

From this formula, we get (15) when substituting the expression found for R.

17. Quantum states and energy levels. Although the atomic model proposed by Bohr was profoundly modified later on, the fundamental postulate of that theory, which is the existence of discrete energy levels, is today an experimentally established fact, not only for the hydrogen atom but for all atoms and molecules, as will be seen in the following sections. We should remember, therefore, that *an atom or a molecule is a mechanical system capable of being in different states* (*quantum states*) *to each of which there corresponds a different value of the energy.* As an intuitive representation of these states, we may (if we wish) think of them as corresponding to the motion of the atomic electrons along different "quantum orbits" (generalization of the privileged circles in the Bohr theory). But such an interpretation should no longer be regarded as realistic.[7] Whatever is said in the following sections is independent of such a model, although it will often be referred to in order to help intuition and to render the language more expressive.

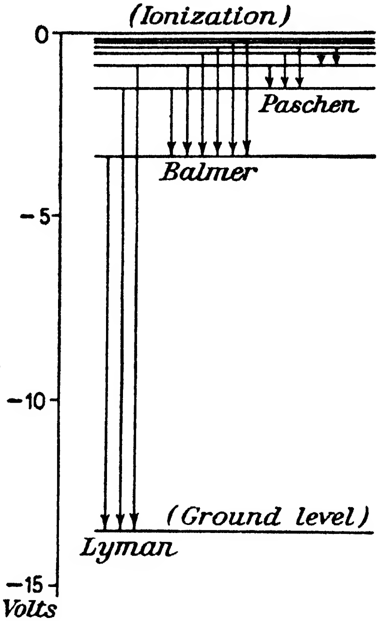

Fig. 8

The different energy levels of an atom (or of a molecule) are usually represented graphically by means of horizontal lines at various heights (see Fig. 8, which represents the levels of the hydrogen atom). Since the energy is always negative, they lie always below the zero level, which corresponds to the ionization limit, toward which they normally converge. The lowest level ($n = 1$) is called "ground level," because it corresponds to the condition in which the atom is normally found (*ground state*); the

---

[7] We can regard such a model as analogous to the well-known hydrodynamic models which aid in the comprehension of the fundamental laws of electricity, by which, for instance, an electric current is likened to the flow of a liquid through a tube.

other states are called *excited states* (in the Bohr model, the ground state corresponds to the smallest orbit).

Another one of the fundamental ideas of Bohr, which still retains its full value, is that an atom can pass from one state to another by absorbing or emitting the energy difference which—when it is in the form of radiation—has a frequency given by law (18). In this manner we interpret the existence of "spectral terms" for any atom or molecule; these correspond to the energy levels by reason of (19), and therefore they permit the energy level scheme to be constructed directly from spectroscopic data alone. In the diagram it is customary to indicate by vertical arrows the transitions that give rise to the observed spectral lines. Their frequency is thus proportional to the length of the arrow.

We can say that the principal purpose of atomic mechanics is to lead to a quantitative prediction of the energy levels of the atoms, which today is done by criteria different from those of Bohr. We shall postpone these considerations until later on, when we shall point out those other experimental facts which confirm the existence of discrete quantum states and the exactness of law (18) in regard to transitions between them, with the emission or absorption of radiation.

18. Excitation and ionization by electronic impact. An atom may acquire energy not only by absorption of radiation but also by collision with another atom or an electron. In terms of the Bohr model, an atomic electron, upon being hit by another particle, may pass from the ground orbit to a more external one (*excitation*), removing the necessary energy from the kinetic energy of the impinging particle.[8] If the impact is sufficiently violent, the electron may be entirely removed from the atom, thus giving rise to the *ionization* of the latter: ionization is thus a limiting case of excitation.

The experiments described in the succeeding sections are concerned precisely with the excitation of atoms by electron impact; therefore we shall anticipate a few general considerations about that phenomenon.

---

[8] Excitation by collision plays a fundamental part in the ordinary temperature emission of radiation. In fact, the latter radiation occurs in the following manner: the collisions due to thermal agitation excite a few of the atoms or molecules (at the expense of kinetic energy or of thermal energy); these atoms (or molecules), when returning to the ground state, emit the energy received in the form of radiation.

First of all we observe that, in view of the very large mass an atom has compared with an electron, the kinetic energy an atom can receive in a collision with an electron is entirely negligible, and we may proceed on the assumption that the atom represents a fixed obstacle for the electron. The energy that is communicated to the atom by the collision serves entirely to increase its internal energy, that is, to induce a transition from the ground state to an excited state. This being understood, the collision may have various effects according to the energy of the colliding electron.

First of all, it is clear that the collision can produce no effect whatsoever, if the energy of the colliding electron is below that required to raise the atom into the *first* of the excited states, that is to say, if it is less than $E_2 - E_1$, which limit is called *resonance energy* for reasons that will become apparent in what follows. In such a case, therefore, the atom remains unaltered, and the electron goes away as if it had collided with a perfectly elastic body; for this reason, such collisions are called *elastic collisions.*

If, on the other hand, the kinetic energy of the colliding electron is even slightly higher than the resonance energy, it may happen that the colliding atom becomes excited, and hence that the electron departs with a lower kinetic energy exactly equal to the excess of the initial kinetic energy above the resonance energy $E_2 - E_1$. It is as if the electron had collided with a soft body: such collisions are called *inelastic.* If the initial kinetic energy is even greater than $E_3 - E_1$, then there may occur, in addition, inelastic impacts that transport the atom into the energy state E_3 and hence remove from the electron an amount of energy $E_3 - E_1$ (*second excitation energy*),[9] and so forth. Finally, if the colliding electron has an energy above $-E_1$ (*ionization energy*), inelastic collisions will occur, resulting in the ionization of the atom, and hence in the appearance of a positive ion, in addition to two free electrons.

When an atom becomes excited, it will generally not remain in that state but after a very short time (which from experiments of Wien turns out to be of the order of 10^{-8} sec) will return to the ground state, emitting the residual energy in the form of radiation, with a frequency given by Bohr's formula (18). Thus we can see

---

[9] Each atom evidently has a series of excitation energies, the first of which is called "resonance energy"; they converge to an upper limit, which is the ionization energy.

that when atoms are bombarded by electrons of energy sufficient to excite them, they must emit radiation. More exactly, if the kinetic energy of the electrons lies above $E_2 - E_1$ (but not above $E_3 - E_1$), the atoms can be excited only to the level E_2 and hence can only emit radiation of frequency $(E_2 - E_1)/h$; the emitted radiation, when analyzed with a spectroscope, will thus yield a single line. If instead the kinetic energy of the electrons exceeds $E_3 - E_1$, some atoms will be excited to state E_2, others to state E_3, and hence the two lines of frequency $(E_2 - E_1)/h$ and $(E_3 - E_1)/h$ can be emitted; sometimes it may even happen that an atom excited to state E_3 will return to the ground state in two steps, that is, first by passing to the state E_2 and from there to state E_1. In this event, we shall see the emission of a line with frequency $(E_3 - E_2)/h$ during the first of these transitions. Thus, if we continue to increase the energy of the electrons, we observe the emission of a more and more complete spectrum. We therefore have, by spectroscopic observation, another method of checking the phenomena of excitation by collision.

In the succeeding sections we shall show briefly how these phenomena have been verified experimentally in a great number of investigations, among which are the classic ones performed by Franck and Hertz in 1913–1914.[10] The fundamental principle of these experiments goes back to an even earlier period, since, starting in 1902, Lenard had applied it in a remarkable series of researches on the ionization by collision.

Before describing these experiments, we should like to make a general observation in order to explain a few concepts that have become customary in this field of study. The arrangement generally used to accelerate electrons to a given velocity consists (see Fig. 9) of a metallic filament, made red-hot by passing a current, thus emitting electrons of low velocity because of the thermionic effect. In front of the filament, a thin metallic net, called a *grid*, is maintained (by means of a battery and a potentiometer) at a *positive* potential with respect to the filament, so that the electrons emitted by the latter find themselves in an electric field which accelerates them toward the grid. They then pass through its meshes and continue beyond because of their inertia. Thus, if V is the potential difference between the filament and the grid, and if we assume that the velocity with which the electrons are expelled from the filament is negligible, we see, by applying the theorem of kinetic energy, that they reach the grid with a kinetic energy equal to eV, and hence with a velocity

---

[10] *Verh. d. D. Phys. Ges.* **16**, 12 (1914).

$$v = \sqrt{\frac{2eV}{m}}. \qquad (22)$$

Thus, by varying V by means of the potentiometer, we can regulate the velocity v at will.

This procedure has given rise to the practice of characterizing the kinetic energy and the velocity of the electrons by indicating directly the potential V, expressed in volts, necessary to produce that energy, and to call this potential the "energy expressed in volts"[11] or, even more improperly, the "velocity expressed in volts."[11] For instance, an electron is said to have a velocity of 1 volt, if it has that velocity which it would acquire by falling through a potential difference of 1 volt, that is, a velocity of 5.932×10^7 cm/sec. We observe that the velocity is really not proportional to the "velocity expressed in volts" but to its square root [see equation (22)].

For the same reason, it has become an established practice to designate the energy of excitation or of ionization of atoms by indicating the potential difference, in volts, necessary to impart that energy to electrons. One thus speaks of "resonance potential," "excitation potential," and "ionization potential"; and energy levels are also often expressed in volts. Given one of these potentials, V, we obtain the corresponding energy E by the obvious relation

$$E = eV, \qquad (22')$$

where e is the charge of the electron in absolute value. From this equation we find that an "energy of 1 volt" is equivalent to 1.602×10^{-12} erg.

Usually the resonance, excitation, and ionization potentials of the atoms are of the order of a few volts, rarely above 20 volts.

Finally we note that to each excitation or ionization potential there corresponds a well-determined frequency (according to Einstein's relation between energy and frequency), given by

$$\nu = \frac{E}{h} = \frac{eV}{h}, \qquad (23)$$

which represents the frequency of a light quantum having the same energy as an electron which has fallen through a potential difference V. If then ν is expressed in cm^{-1} (which will be indicated by $\tilde{\nu}$ as usual), equation (23) becomes

$$\tilde{\nu} = \frac{eV}{hc}. \qquad (23')$$

Frequently the energy levels of atoms are simply expressed in cm^{-1} according to (23′); 1 volt corresponds to 8067.9 cm^{-1}.

---

[11] Often we say "in electron-volts."

If, on the other hand, the radiation is characterized by the wavelength λ, the relation (23′) must be replaced by

$$\lambda = \frac{hc}{eV}. \tag{24}$$

From (24), upon substituting the numerical values, we obtain the following easily remembered relation between λ and V, where λ is expressed in angstrom units and V is expressed in volts:

$$\lambda V = 12{,}395. \tag{24'}$$

19. Experiments of Franck and Hertz. The experiments of Franck and Hertz consist in the excitation of atoms by electronic impact and in the indirect measurement of the kinetic energy lost by the electrons in the collisions, or else of the energy imparted to the atom when it is being excited. In this fashion we obtain the heights of the various energy levels above the ground level, or else the "excitation energies" $E_2 - E_1$, $E_3 - E_1$, and so on. Comparison of the values found with the values obtained from spectral terms constitutes a verification of the fundamental Bohr hypothesis expressed by (18).

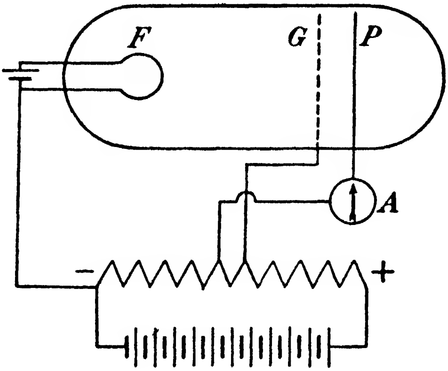

Fig. 9

The arrangement of Franck and Hertz is schematically the following (Fig. 9). In a glass tube containing the substance under study in a rather rarefied gas or vapor state, the system for the acceleration of the electrons is located. As has been stated in the preceding section, this system is composed of an incandescent filament F and of a *grid* G, maintained at a positive potential with respect to the filament F by a battery. A potentiometer permits this potential difference to be regulated, and a voltmeter (not shown in the figure) indicates it exactly. The field existing between F and G accelerates the electrons, which rush toward the metallic net and partly traverse it. On the other side of the grid, at a distance of about 1 mm, there is a *plate* P, maintained at a potential slightly below that of the grid, so that the electrons are slightly slowed down in that space. Having reached the plate P, they return to the filament through a galvanometer A.

If the electrons suffer no inelastic collisions, they will move with accelerated motion up to the grid, where they will attain their maximum velocity. This maximum corresponds to a kinetic energy equal to eV (if V is the potential difference between filament and grid). This kinetic energy is more than enough to make the electrons overcome the weak opposing field existing between grid and plate, so that those which pass between the meshes of the grid all arrive at the plate and are registered by the galvanometer.

But now let us suppose that we increase the potential V until eV is slightly larger than $E_2 - E_1$: the kinetic energy of the electrons in the neighborhood of the grid will then be sufficient to produce inelastic impacts; some of the electrons will lose almost all their kinetic energy in a collision and will no longer be able to overcome the opposing field and to reach the plate. Hence the galvanometer, as soon as eV exceeds the limit $E_2 - E_1$, will show a sudden decrease in current. Reading the potential, V', for which this discontinuity occurs, we can calculate the first excitation energy (or resonance energy) by the formula

$$E_2 - E_1 = eV'.$$

This V' is thus the potential we have called *resonance potential.*

If the potential is increased further, the current also increases, until a new sudden diminution occurs for a value V'' nearly twice that of V'. Indeed, when V is raised, it follows that the inelastic collisions, rather than being produced only in the vicinity of the grid, are already produced ahead of it (that is, closer to the filament.) The more V is raised, the more the region recedes in which such collisions occur, so that the electrons which have lost all their velocity in a collision can still be accelerated by the field before arriving at the grid; when V reaches $2(E_2 - E_1)$, some of these electrons, after the first inelastic impact, reacquire in the field an energy $E_2 - E_1$ and hence undergo a second inelastic collision near the grid, in which they once again lose their kinetic energy and thus remain incapable of reaching the plate. Similarly it is evident that there will be another dip in the current for $V = 3(E_2 - E_1)$, and so on. In fact, the experimental curves exhibit the behavior represented by Fig. 10, which applies to mercury vapor (Einsporn).

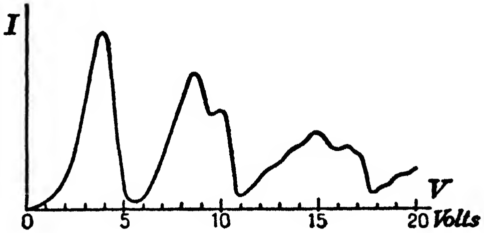

Fig. 10

We may take advantage of this fact by measuring with great accuracy the resonance potential V', obtaining it from the distance between two consecutive maxima of the curve, rather than from the position of the first maximum; in this manner we eliminate those causes of error (contact potentials, for example) which produce a displacement of the curve as a whole or which perturb its first part.

Thus far we have reasoned as if the atom could absorb only the energy $E_2 - E_1$, thus omitting the possible transitions from the ground level E_1 to the levels E_3, E_4, and so on. This omission is permissible in the many cases where collisions are so frequent that almost all electrons, as soon as they have reached an energy sufficient to carry an atom to the level E_2, lose this energy in a collision and thus are no longer able to produce an excitation requiring a higher energy.

However, by varying the conditions of the experiment, this effect can be prevented (it is necessary to decrease the pressure of the gas); then different series of maxima occur in the curve, each series corresponding to an excitation level.

Finally, we point out that the same apparatus permits the direct determination of the ionization potential: all we need do is to establish a considerable and fixed potential difference between grid and plate, larger than any potential used to accelerate the electrons (for example, 40 volts), instead of a potential difference small compared with the potential to be determined. None of the electrons projected from the filament will then be able to reach the plate P, since they will all be driven back by the field existing between G and P. But if some of the colliding atoms become ionized in the vicinity of the grid, they will remain positively charged and thus will rush toward the plate and be registered by the galvanometer. It will then suffice to increase the accelerating potential of the electrons gradually, until the galvanometer begins to read a current; this potential V_i will be the one necessary to ionize the gas, which amounts to saying that

$$eV_i = -E_1.$$

This cursory description of these experiments will be sufficient to give an understanding of their fundamental concept and to show the possibility of determining the various energy levels by electrical means. We cannot describe here all the particular techniques and the many variations of these experiments which were performed on various substances, not only by Franck and Hertz but by many other physicists; for these data the reader is referred to No. 28 of the Bibliography.

20. Optical verification of the excitation by collision. In the experiments described in the preceding section, the energy absorbed by the colliding atom is determined indirectly. A more direct method of measuring it consists, as has been mentioned, in observing the radiation emitted by the atom when it is returning to its ground state.

The apparatus consists essentially of a tube of quartz glass containing the usual incandescent filament and grid, between which there is maintained a known accelerating voltage. Then an image of the region where the collisions occur is projected upon the slit of a spectroscope, and the investigator determines those values of the voltage for which the various lines of the spectrum appear. This method was used with mercury vapor by Franck and Hertz in 1914. They found that the line $\lambda = 2536.6$ Å (*resonance line*) of that gas did not occur so long as the potential difference was below 4 volts but that it already was susceptible of being photographed at 5 volts. More precise measurements have yielded 4.9 volts for the resonance potential of mercury. From the wavelength of the resonance line, we find, by means of formula (24′), that the potential necessary to excite this line should be 4.87 volts, in perfect agreement with the value found experimentally.

Analogous experiments have been repeated by many other observers and on many different substances, always with results in full accord with the theory (see No. 28 of the Bibliography).

It should be noted that very often the radiations obtained by electronic impact lie in the ultraviolet region. In that case it is possible to detect them not only photographically but also by making use of their photoelectric effect. If we gradually increase the accelerating voltage of the electrons, following simultaneously the variations of the photoelectric current, and if we represent this dependence graphically, we observe that the curve suffers sudden deflections corresponding to the excitation potentials, since to each of these potentials there corresponds the emission of new lines and hence an increase in the photoelectric effect. This was the method used by Horton and Davies, Franck and Knipping, Foote and Mohler, and others.

All these experiments, while establishing a strict connection between optical phenomena and the phenomena of electronic collisions, constitute a direct proof of the physical reality of energy levels, as well as of the validity of the Bohr formula relating the emitted frequency to the energy difference of two quantum states.

21. Collisions of the second kind. As we have seen, in a collision between an electron and an atom it may happen that the electron gives up part of its kinetic energy to the atom in the form of excitation energy. Klein and Rosseland have observed that in addition to this type of collision, said to be *of the first kind*, there must exist the possibility of collisions representing the inverse phenomenon; that is to say, an already excited atom colliding with an electron gets rid of its excitation, not by radiating but by giving its excitation energy to the electron in the form of kinetic energy. Hence the electron recoils with a velocity higher than it had before. The necessity for the existence of these collisions, called *collisions of the second kind*, has been demonstrated by Klein and Rosseland by means of the following thermodynamic argument.

Let us suppose that a gas formed of atoms and electrons is enclosed in a cavity with walls that are perfectly reflecting for atoms as well as for electrons and radiation, so that no energy exchanges with the outside are possible. As a consequence of the collisions of the first kind, the electrons will gradually give up their kinetic energy to the atoms. It there were no collisions of the second kind, the average kinetic energy of the electrons would continually diminish with respect to the energy of the atoms. In this way two mixed gases would result, which, rather than tending toward a temperature equilibrium, would get further and further away from it; this condition would contradict the second law of thermodynamics.

The existence of collisions of the second kind is thus a thermodynamic consequence of the experimentally observed collisions of the first kind.

The collisions of the second kind also have macroscopic effects of great practical importance. In fact, it is by virtue of these collisions that a gas traversed by radiation heats up, since the radiation absorbed by the molecules excites them and then is converted into molecular kinetic energy (that is, into heat) by means of the collisions of the second kind.

22. Optical resonance. The scheme of the energy levels lends itself to the interpretation of the phenomena of optical resonance and of fluorescence, as well as those of ordinary emission and absorption. However, as we shall see, the phenomenon of so-called optical resonance can be interpreted equally well by the classical scheme.

This phenomenon, discovered by Wood, consists of the following: if a metallic vapor is irradiated with light of a wavelength exactly corresponding to one of its definite absorption lines, that light is not genuinely "absorbed" by the gas (that is, transformed into heat) but is *scattered* almost entirely in all directions. Characteristic of this phenomenon are the surprisingly large intensity scattered even by a very rarefied vapor, and the almost absolute monochromaticity of the scattered light.

The experiment, performed for the first time with mercury vapor, is done in the following manner. Into a large, well-evacuated quartz bulb is placed a drop of mercury, so that the bulb contains mercury vapor saturated at ordinary temperatures, and thus of very low density. If this bulb is exposed to the rays of a quartz-walled mercury lamp, which strongly emits the ultraviolet line of wavelength $\lambda = 2536.6$ Å, it will be found that the bulb intensely re-emits that light, which can be photographically detected. Spectrographic analysis of the light of the lamp after it has traversed the bulb shows that the line λ2536.6 has been strongly absorbed. The phenomenon does not take place if the incident light has a wavelength ever so slightly different from this value, and accordingly such a line is called a *resonance line.*

The phenomenon can also be observed with a bulb filled with sodium vapor and illuminated by the yellow *D* line of sodium (or else by one of its components, $\lambda = 5890$ Å and 5896 Å). In this case there is the advantage that the light is visible without photography, and the bulb appears brightly radiant to the eye; however, it is necessary to heat the bulb, because the density of saturated sodium vapor at ordinary temperatures is too small.

The classical interpretation of resonance, from which the name of the phenomenon is derived, is as follows. The atoms of the vapor contain oscillators having frequencies equal to those contained in the emitted light. If we illuminate the atoms with light of one of these frequencies, the corresponding oscillators, finding themselves excited by an alternating electric field with a frequency equal to their own, will go into vibration of rather large amplitude, because of the phenomenon of "mechanical resonance," well known also in acoustics; hence the oscillators will emit intense light in all directions.

In contrast, the quantum theory interprets the phenomenon of resonance as follows. The atoms finding themselves in the ground state can pass to a higher level, not only under the action of an electron collision but also under the action of incident radiation, that is, of photons; provided, however, that each of the photons contains exactly the energy necessary for the excitation of the atom. Thus, if the atom is struck by light of frequency

$$\nu = \frac{E_2 - E_1}{h},$$

it can go from state 1 to state 2. However, it normally does not remain in the second state, as we have already pointed out, but returns to the ground state by performing the inverse jump, and hence emits the absorbed energy in the form of radiation of precisely the same frequency.

This theory makes possible the explanation of many peculiarities of the resonance phenomenon which remain unexplained in the classical theory.

23. Fluorescence. A phenomenon related to resonance but considerably more common and well known for a long time, is that of fluorescence. In fluorescence certain substances, when illuminated, re-emit light in all directions, of spectral composition different from that of the incident light. For instance, a solution of fluorspar, of extract from the bark of the horse chestnut (aesculin), or of sulfate of quinine, when placed in a glass container and illuminated with white light, appears (when viewed laterally) radiant with azure or greenish light. A chlorophyll solution, on the other hand, presents a red fluorescence.

The phenomenon of fluorescence generally obeys the following law, discovered empirically by Stokes, whose name it carries: *fluorescent light does not contain frequencies above that of the exciting light (which we assume to be monochromatic).*

This phenomenon may be explained by the quantum theory in the following manner. As we have said, the atoms become excited by the incident light; if they are carried from the ground level E_1 to the one immediately above, E_2, we are dealing with resonance, of which we have already spoken. If instead they are carried into a higher energy level, such as E_3, they may then, in order to return to the ground state, perform two (or more) successive quantum jumps, in each of which they emit light of a different frequency. For instance, from the level E_3 they would go to the level E_2 with the emission of the line of frequency $(E_3 - E_2)/h$, thence passing from E_2 to E_1 with emission of frequency $(E_2 - E_1)/h$. Similarly, if the excitation had carried the atom to an even higher level, it could descend by different successive jumps. Hence it can be understood how a single exciting frequency may give rise to the emission of different frequencies. In this way Stokes's law is immediately justified, since it is evident that all the quantum jumps performed during the emission phase correspond to level differences smaller than that corresponding to the excitation, and hence to frequencies lower than that of the exciting light.

24. Sensitized fluorescence. A significant confirmation of the quantum theory of fluorescence is furnished by the following experiment, performed for the first time by Cario[12] and further studied by Franck and Cario and by others. A quartz tube, at rather elevated temperatures, contains a mixture of mercury vapor and the vapor

---

[12] *Zeits. f. Physik* **10**, 185 (1922).

of another metal, thallium for example. The tube is illuminated with light of wavelength 2536.6 A emitted by a mercury arc, and the experimenter observes the laterally scattered light with a spectroscope. He finds, in addition to the resonance line of the mercury, a certain number of lines characteristic of thallium. However, these do not appear if he illuminates some thallium vapor unmixed with mercury vapor, or if the incident light has a frequency different from the resonance line of mercury.

This phenomenon, called *sensitized fluorescence* or *indirect fluorescence*, forces us to think that the state of excitation induced by the light in one of the gases may be transmitted to the other gas, and is interpreted as follows. Many mercury atoms absorb the incident radiation and go to an excited state. If they happen to collide with a thallium atom while they are still in that state, they may get rid of their excitation energy by transferring it to a thallium atom and thus carrying it into an excited state. The thallium atom in turn, returning in one or more jumps to the normal state, emits its characteristic lines.

Of course, the energy furnished by the excited mercury atom is not equal to the energy necessary to excite the thallium atom. It is normally superior to the latter, and the surplus energy is imparted to the thallium atom in the form of kinetic energy.

25. Note on the development of spectroscopy. The spinning electron. In general it is not possible to calculate quantitatively the energy levels of an atom with several electrons by the Bohr-Sommerfeld theory, as has been possible in the case of hydrogen. In fact, already in the case of only two electrons (helium atom) the problem presents serious mathematical difficulties. It has been treated with approximations, and we shall speak of the results of these laborious calculations shortly. In the more complex cases the mathematical difficulties become practically insurmountable. Nevertheless it has been possible to attack the problem, at least qualitatively, thanks to a simplifying hypothesis proposed by Bohr, which has proved to be in good agreement with reality. This hypothesis consists in the assumption that the quantum jumps which give rise to radiation generally involve only one electron—the most easily excited electron or else the electron whose normal energy level lies highest. This electron is called the *emission electron;* it is assumed that the others continue to describe their orbits around the nucleus undisturbed, forming with the nucleus an invariant structure called the *core* or the *remainder of the atom.* Thus the problem is reduced to a study of the energy of the quantum states of the emission electron; it differs from the analogous problem for hydrogen, inasmuch as the emission electron is subject to the force exerted by the whole remainder of the atom, not only by the nucleus. This force is of course quite complex and varies in time. In general we must limit ourselves to a rough evaluation by substituting for it a time-independent central field. The problem is thus reduced to one of central motion under the action of a non-Newtonian force. Solving it and then applying the three Sommerfeld conditions (the system having three degrees of freedom), we succeed in finding for the emission electron a series of energy levels (depending upon three quantum numbers[13]) which are in general in good qualitative agreement with observed values. However, we are far from being able to attain quantitatively exact results.

Such considerations have served as the basis for an enormous amount of interpretation and organization of experimental results in the field of spectroscopy. This work consists essentially in assigning to each spectral term a group of two quantum numbers n and l. This assignment retains its value independently of the mechanical model of Bohr and Sommerfeld, and it still constitutes the key for the interpretation of spectra. In fact, a series of rules, discovered partly empirically, partly by theoretical considerations (and today for the largest part justified by quantum mechanics), enables us to deduce many of the characteristics of a line, such as intensity, Zeeman effect, Stark effect, and others, by using the quantum numbers corresponding to the two spectral terms giving rise to the line.[14]

But from this search for quantum numbers corresponding to each energy level there arose the necessity for a new hypothesis. It was found that three quantum numbers (or rather two in the absence of a magnetic field) are insufficient to define an energy level, and a satisfactory scheme was obtained only when a fourth quantum number, called *inner quantum number,* was introduced, which, how-

---

[13] They are called: *principal quantum number n, azimuthal quantum number l, magnetic quantum number m.* The last does not influence the energy, except in the case of the Zeeman effect; normally, however, the terms depend only on the first two.

[14] See, for instance, No. 23 of the Bibliography.

ever, could only assume two values in each case. There remained the problem of its mechanical interpretation. After a hypothesis of Landé had proved insufficient for the purpose, a hypothesis stated in 1925 by Uhlenbeck and Goudsmit,[15] and independently by Bichowsky and Urey,[16] proved to be entirely satisfactory. It has been of decisive importance in the further development of atomic physics, proving to be an ever-growing source of results agreeing with experience. This hypothesis consists in attributing to the electron, in addition to electric charge and mass, a magnetic moment of magnitude equal to $eh/4\pi mc$ (that is, to a *Bohr magneton;* see §61, Part II), and a (mechanical) angular momentum having the same direction and having a value $h/4\pi$, as if we were dealing with a small top whose axis was magnetized. At one time it was thought that this property could be interpreted by imagining the electron as a charged sphere rotating about an axis: the rotation of the mass was to have produced the mechanical angular momentum; whereas the rotation of the electric charge, being equivalent to a system of circular currents, was to have given rise to the magnetic moment. This model had soon to be abandoned, but nevertheless the hypothesis of Uhlenbeck and Goudsmit has retained the improper name *hypothesis of the spinning electron,* and the angular momentum of an electron is today designated by the name of *spin.* The electron has thus lost its spherical symmetry by acquiring a privileged axis, namely, the axis of its spin.

Applying to the angular momentum of the electron a condition analogous to those of Sommerfeld we are led to assume that when the electron finds itself in a magnetic field, it can take on only two orientations, namely, with its spin parallel or antiparallel to the magnetic field, to each of which there corresponds a different value of the energy. The two values of the inner quantum number are used merely to distinguish between the two energy levels corresponding to these different possibilities of orientation.

The hypothesis of the spinning electron was later shown to be very adaptable to the interpretation of the magnetic properties of metals; in fact, some phenomena of this class (gyromagnetic effects) have made it possible to measure the ratio of the magnetic moment to the angular momentum of the electron, which yielded results in agreement with the values given above.

---

[15] *Nature* **117**, 264 (1926); *Physica* **5**, 266 (1925).

[16] *Proc. Nat. Acad. Sci.* **12**, 80 (1926).

Finally we shall mention the fact that the methods of interpretation of atomic spectra have been successfully extended to the spectra emitted by molecules (*band spectra*), many of which, especially in diatomic molecules, have been perfectly interpreted by means of the same theoretical principles. We shall, however, confine ourselves here to the consideration of the mechanics of atoms, referring to other volumes for the molecules.

26. The crisis of the modelistic theories. Although the Bohr-Sommerfeld theory served as a guide for the theoretical and experimental investigations mentioned in the preceding sections and for many others, it nevertheless rests upon postulates which are so unsatisfactory (mainly because of the arbitrary introduction of discontinuity) that it is no longer considered to be a definitive expression of a physical theory. Instead it is viewed as a provisional codification, so to speak, of the changes that must be made in classical mechanics and electromagnetic theory in order to make them applicable to the atomic domain. Furthermore, although the general concept of energy level and the Bohr formula (18) concerning frequencies have proved in all cases to be in such close agreement with experimental results as to make us feel certain that they represent actual physical reality, the experimental proofs of the exactness of the method of determination of levels by calculation of the mechanical orbits and the successive application of the Sommerfeld conditions are considerably less solid. In fact, the only case in which this method leads to quantitatively exact results is the one of hydrogenlike systems. In the case of helium and in the few other cases in which the method has been applied (the serious mathematical difficulties being overcome by approximations), energy levels have been found in open contradiction with experiment. Later it was found in certain cases (for example, in the prediction of the anomalous Zeeman effect) that the formulas deduced from the Bohr-Sommerfeld theory require slight modifications, such as the substitution of $l(l+1)$ for l^2, where l is a quantum number. These modifications were determined empirically without any justification for them being found within the realm of the Bohr-Sommerfeld theory.

This theory then encountered other serious difficulties in its application to optical dispersion, to certain features of collisions between electrons and atoms (Ramsauer effect), and to various other questions.

While these difficulties accumulated in the field of atomic mechanics, the situation in the field of the theory of light appeared no less serious: as we have mentioned, there existed two models, one undulatory and one corpuscular, each of which permitted the exact interpretation of one category of phenomena, but was incompatible with the other.

The various attempts to substitute more coherent representations for the atomic model of Bohr and Sommerfeld and for the two models for light were all doomed to failure, so that the situation of theoretical physics around 1925 was characterized by a sense of discomfort and by the feeling that a profound revision of the fundamental principles of theoretical physics was necessary.

The new logical form which these principles were to take appeared clearly only in the following year, when the new quantum mechanics had already achieved brilliant successes on a purely formal basis. But already in 1925 several of the major exponents of theoretical physics had formed the conviction that this critical situation had arisen from the presupposition, inherent in all previous theories, that the atoms, photons, and other entities of the atomic domain must be thought of in terms of extremely small bodies or mechanisms, even though subject to laws entirely different from those governing ordinary bodies. They may, however, be essentially different entities, to which our intuitive concepts of body, motion, and so on, do not apply (indeed, there is no reason why they should apply). In other words, we must consider the possibility that the atomic world, though governed by mathematically expressible laws, is not representable by any intuitive model (see §34).

The new direction which theoretical physics came to take after 1925 as a consequence of these new ideas is usually designated by the name of *quantum mechanics*.[17] It has been developed in different forms, of which we shall give a purely historical and informative account in the following chapter, reserving for other parts of the book the systematic presentation of their essential features.

---

[17] This expression is not to be confused with the nowadays rather generic one of *quantum theory*, which embraces all the theories in which Planck's constant h has an essential part. Quantum theory thus includes the Bohr-Sommerfeld theory (now sometimes called *classical quantum theory*), as well as *quantum mechanics* in its various forms (*wave mechanics*, *matrix representation*, and *operator representation*).

CHAPTER 4

Quantum Mechanics

27. The quantum mechanics of Heisenberg (matrix method). The new approach mentioned at the end of the preceding chapter was inaugurated by W. Heisenberg in a paper[1] published in July 1925. The fundamental idea advanced in that work is that some of the quantities inherent in the atomic model used in the quantum theory (such as the coordinates of an electron within an atom at a given instant or the duration of an orbital revolution) are no longer being measured directly. Hence, in view of the fact that reasoning based upon these quantities leads to known difficulties, it is legitimate to doubt their actual physical significance and the possibility of their future measurability. Other quantities, however (such as emitted frequencies, intensities, and so on), are directly observable. Hence instead of looking for a geometrical and mechanical model that will make it possible to obtain values of observable quantities from a nonobservable structure, it is better to try to relate the observable quantities directly, without the intervention of any model.

However, the direct relations between observables are in general not expressible by ordinary algebra, and hence the further development of Heisenberg's idea led to the utilization of *matrix algebra*, a mathematical process which had been known for some time but which had not yet found any applications in the field of physics. This method was greatly developed, mainly by Heisenberg, Born, and Jordan. In many cases it led once again to the results of the Bohr-Sommerfeld theory, whereas in others it led to results which are in even better accord with experience.

The importance of the progress achieved by Heisenberg is not limited to these considerations. He was also able to deduce all the results, by a unified method, from the same organic system of postulates, and hence to substitute, for a theory based on partially

---

[1] *Zeits. f. Physik* **33**, 879 (1925).

contradictory foundations, a theory perfectly coherent from a logical point of view. However, the matrix method contains within itself, along with these merits, the serious inconvenience of being rather difficult to comprehend and of not satisfying the needs of intuition. These objections arise from the use of such an unusual and somewhat complex mathematical apparatus as matrix algebra, but above all from the required renunciation of any geometrical or mechanical model. This renunciation proves to be necessary in order to be able to formulate in a precise and coherent form the laws of the atomic world.

The intimate reason for the impossibility of basing atomic mechanics on a mechanical model without sacrificing logical coherence or precision has been elucidated in a later paper of Heisenberg entitled "Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik" ("On the Intuitive Content of Quantum-theoretical Kinematics and Mechanics"),[2] in which the so-called "uncertainty principle" is established. This principle, which may be termed the key to all atomic mechanics, has made it possible to place quantum mechanics in its true light. We shall return to this principle and the consequent impossibility of a rigorous atomic model in greater detail in what follows.

28. Wave mechanics. Almost simultaneously with the matrix method, another method for treating atomic problems arose and was developed. This theory, called *wave mechanics*, was originally suggested by L. de Broglie.[3] Later it was developed and placed in a new perspective by E. Schrödinger, who outlined its fundamentals in a series of lucid papers published from February 1926 on.[4]

The Schrödinger method starts with the observation (which goes back to Hamilton) that the classical laws of point mechanics can be put into a form analogous to the laws of geometrical optics (for example, the principle of least action for the motion of a point is analogous to Fermat's principle for light rays). Some of these analogies can even be seen without having recourse to formulas. For instance, in a homogeneous medium, light is propagated with rectilinear uniform motion, just like a point particle in a field of

[2] *Zeits. f. Physik* **43**, 172 (1927).

[3] Thesis presented to the University of Paris, 1924; *Ann. de Physique* **10**, 3, 22 (1925).

[4] *Ann. d. Physik* **79**, 361, 489, 734 (1926). These and other fundamental papers are collected in one volume (see Nos. 17 and 17b of the Bibliography).

uniform potential. If instead the index of refraction varies from point to point, the light rays will be curved (such as in a mirage), and their velocity is no longer uniform. This is similar to the motion of a point particle if the potential is a function of position (force field). The potential in this *analogy* corresponds to the index of refraction, as we shall see.

Now it is known that the laws of geometrical optics fail in all phenomena (diffraction) where screens, slits, bars, and other obstructions of dimensions very small or at least comparable with a wavelength are involved. In these cases we must employ wave optics instead, of which geometrical optics represents only a first approximation. Similarly, according to Schrödinger, classical point mechanics represents only a first approximation of more general mechanical laws. If we want to retain the parallelism with optics, we must consider that these laws are analogous to the laws of wave optics. There must then exist in mechanical problems a quantity corresponding to the wavelength; and when dealing with systems of dimensions large compared with that quantity, we shall be able to use classical mechanics with sufficient approximation. However, in systems of dimensions comparable with a wavelength, phenomena corresponding to those called *diffraction* in optics will occur, and hence it will be necessary to use wave mechanics. Therefore if the mechanical wavelength is of atomic dimensions (that is, of the order of magnitude of 1 angstrom), there may be an explanation of how classical mechanics applies well to ordinary bodies but fails in the interpretation of atomic phenomena. Schrödinger's idea, then, consisted in trying to construct a mechanics which bears the same relation to classical mechanics as does wave mechanics to geometrical optics.

This theory is developed in rather singular fashion, inasmuch as it has at first the appearance of a purely mathematical operation applied to an abstract quantity ψ, satisfying an equation characteristic of wave phenomena. Later Schrödinger sought to relate this quantity to a physical model by interpreting it as an expression of the electric charge density. It was only later that Born[5] suggested its probabilistic interpretation, which today, in the light of the uncertainty principle, must be recognized as the only legitimate one.

---

[5] *Zeits. f. Physik* **37**, 863; **38**, 803 (1926), and **40**, 167 (1927).

The Schrödinger theory, even in its preliminary phase, in which it represented a mathematical procedure of unknown and doubtful meaning, immediately gave the impression of constituting an enormous progress over the Bohr-Sommerfeld theory and of coming rather close to the profound nature of things. This assurance depended on the fact that wave mechanics does not postulate the existence of privileged orbits and hence of discrete energy levels; the latter emerge as a consequence of a postulate that is more in accord with the normal convention that a certain function be finite and continuous everywhere.

In this theory, the discontinuity arises in an altogether natural manner from the mathematical procedure, in a way quite similar to the procedure by which, in acoustics, it is shown that a vibrating system can produce only discrete notes.

With these principles, Schrödinger was able to calculate the hydrogen spectrum, the Zeeman and Stark effects, the oscillator, and so on, always obtaining results in agreement with those obtained by Heisenberg's matrix method. This coincidence is not accidental, since Schrödinger himself soon showed that the two methods, although originating from very different concepts, are strictly equivalent. That is, they represent two different forms of the same mathematical procedure and must therefore lead to identical results in all cases. Hence in the treatment of various problems we may choose one or the other method, according to circumstances. However, for a general presentation, the Schrödinger method has the advantage of leaning more strongly upon intuition and of requiring a less exceptional mathematical apparatus; therefore we shall use it as an introduction to the more general methods of quantum mechanics.

29. Diffraction of electrons and of atoms. According to wave mechanics, the behavior of a beam of electrons (for example, a beam of cathode rays), atoms, or molecules is governed by mathematical laws quite similar to those which govern the propagation of a bundle of waves—as if the particles were *guided* by the latter. These waves, which we shall call *de Broglie waves,* have a wavelength λ which depends on the velocity v of the particles and which is given, as we shall see in §33, by h/p, where p is the momentum (a relation analogous to the one which holds for photons; see §8). For velocities small with respect to c this formula becomes

$$\lambda = \frac{h}{mv}, \tag{26}$$

where m is the rest mass of a particle. In particular, it can be deduced that if the beam encounters a grating, it must be diffracted like a light beam. But since it is practically impossible, for experimental reasons, to obtain a well-defined beam of very slow particles, the wavelength actually cannot be made to exceed a certain limit, which is of the order of several angstroms (corresponding to velocities of a few tens of volts). Therefore a parallel must be drawn to X rays rather than to light waves, properly speaking; and hence for diffraction we must normally use a crystal instead of an optical grating, in analogy with the experiments of Laue and of Bragg.[6]

Apparently this idea was first suggested by Elsässer, in 1925, soon after the appearance of the first works on wave mechanics. But the experiment on electron diffraction by a crystal was performed for the first time by Davisson and Germer, who made it famous in 1927 and who succeeded thus in measuring the "electron wavelength" as a function of the velocity, finding formula (26) to be fully verified. This discovery, which profoundly altered the usual concepts of the nature of the electron, constitutes the most direct and most remarkable confirmation of Schrödinger's wave theory. The experiment of Davisson and Germer was further modified and perfected by many other investigators, until today there exists a technique, in a certain sense parallel to the spectral analysis of X rays. In a few cases it has even been found advantageous to employ the diffraction of electrons of known wavelength for a study of various problems of structure, as has been done for some time with X rays on a very large scale. In view of the smaller penetration of electrons, the new method is particularly well adapted to the study of the surface layers of the diffracting body.

It has furthermore been verified that diffraction phenomena arise not only from electrons but also from all material particles directed in a beam upon a crystal. This assumption was proved experimentally by Stern and his collaborators and by Johnson and others, who used molecular beams of hydrogen molecules, atomic

---

[6] Nevertheless, Rupp [*Zeits. f. Physik* **52**, 8 (1928)] has succeeded in obtaining the diffraction of electrons from 70 to 300 volts by an optical grating, used with extremely grazing incidence, and to verify the predictions of diffraction theory within 2 or 3 per cent.

hydrogen, and helium. For these particles as well, the validity of formula (26) has been confirmed.[7]

30. Experiments performed by the Laue method. We shall now outline the first memorable experiments of Davisson and Germer.[8]

The electrons were emitted (Fig. 11) from a filament of incandescent tungsten F, and were accelerated by a field established between F and the diaphragm D by battery B. By varying the potential of the latter, it was possible to vary the velocity of the electrons (in the experiments described, the potential was varied from 30 to 370 volts). The diaphragms D and D' served to define a narrow beam of electrons incident normally upon R, a nickel crystal cut parallel to one face of the octahedron.[9] The electrons were reflected by the crystal, and their distribution in various directions was studied by collecting them in a Faraday cup C (protected by a screen against electrostatic influences and provided with a narrow aperture so as to collect only the electrons scattered in a given direction), which was connected to a galvanometer g. By dis-

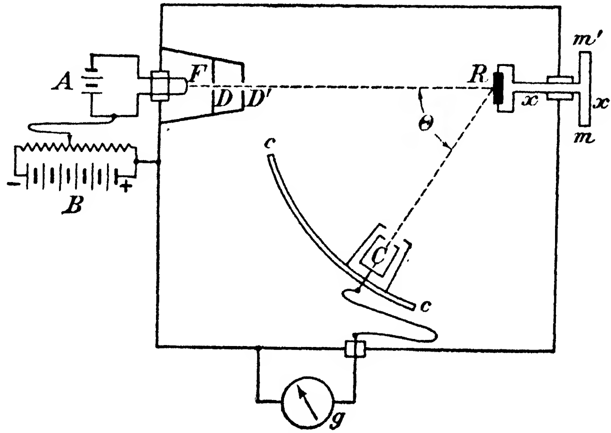

Fig. 11

---

[7] For particulars on the electron and molecular diffraction experiments, and for many references, see Bibliography No. 30 and also the monograph of J. J. Trillat, *Les preuves experimentales de la mécanique ondulatoire* (Paris: Hermann, 1934).

[8] *Phys. Rev.* **30**, 705 (1927); *J. Chem. Ed.* **5**, 1041 (1928).

[9] Nickel crystallizes in a monometric system, and the crystal lattice is of the face-centered cubic type. The faces of the octahedron are designated by (111).

placing the collector C along the graduated arc cc' and observing the deflections of the galvanometer for the various positions, we can study the distribution of the diffracted electrons at various values of the angle θ. Then in order to study the distribution at different azimuths, it was possible to rotate the crystal about the axis xx' and to determine its orientation by means of the graduated circle mm'. Of course, the whole apparatus was enclosed in a highly evacuated vessel.

This experimental arrangement is analogous with the one of Laue for X rays (the only difference being that the diffracted beams were observed on the same side of the crystal instead of on the other side), and its theory can be established on the same basis that is used for the interpretation of the Laue experiment. However, there is one important difference between the two cases. In the electron case it is necessary to take into account the fact that the wavelength inside the crystal has a value λ' different from the value λ that it has outside, because of the different propagation velocity of the de Broglie waves within the material and *in vacuo*. This difference, however, is negligible in the case of X rays, for which the index of refraction for all substances is practically equal to unity. The necessity of taking this fact into account was pointed out by Bethe, who showed how the experiments of Davisson and Germer, when interpreted correctly, were in full accord with the Schrödinger theory. The existence of an index of refraction, in addition to being foreseen for theoretical reasons, is thus demonstrated experimentally; and the experiments of Davisson and Germer also permit us to calculate the value which we shall call μ and which comes out larger than 1, and tends toward 1 with an increase of v. Therefore, when establishing the theory of these experiments, we shall remember that inside the crystal the wavelength will not be λ but

$$\lambda' = \frac{\lambda}{\mu}. \tag{27}$$

We shall now consider a beam of electrons incident normally upon the surface ss of the crystal (see Fig. 12). Let us select one of the planes which may be drawn through the crystal, forming, for instance, an angle φ with the surface. It is known from Bragg's theory that this and all other planes parallel to it behave like partially reflecting surfaces. In general, the plane waves reflected by them will interfere destructively, except for the case where the wavelength λ' has a value such that Bragg's relation is satisfied:

$$n\lambda' = 2d \cos \varphi, \tag{28}$$

where d is the distance between two adjacent crystal planes and n is an integer, or else [see (27)] for the wavelengths λ such that

$$n\lambda = 2d\mu \cos \varphi. \tag{29}$$

In this case we have *selective reflection.*

Let us suppose then that λ has one of those values for which reflection occurs, and let us look for the direction in which the reflected waves will travel outside the crystal. Upon leaving the crystal they suffer a refraction, and hence the emerging ray (normal to the emerging waves) forms an angle θ with the normal to the surface given by

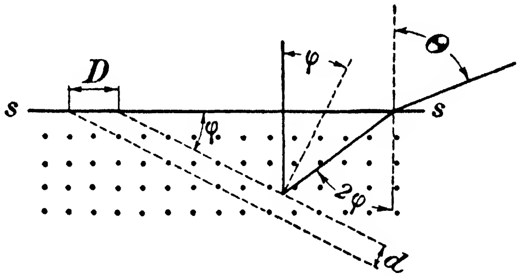

Fig. 12

$$\frac{\sin \theta}{\sin 2\varphi} = \mu,$$

since the angle of incidence upon the surface is 2φ, as can be seen from the figure.

Hence we have $\qquad \sin \theta = 2\mu \sin \varphi \cos \varphi.$

Solving for $\cos \varphi$ and substituting into (29), we get

$$n\lambda = d\,\frac{\sin \theta}{\sin \varphi}. \tag{30}$$

Calling D the distance between the intersections of the two crystal planes with the surface ss (which is a constant of the crystal, known from X ray studies), we get, as seen from the figure,

$$d = D \sin \varphi,$$

and hence (30) becomes

$$n\lambda = D \sin \theta. \tag{31}$$

μ does not enter in this equation, which is the same that holds for X rays. Thus it is possible, from a measurement of θ, to obtain λ without making any assumption concerning the index of refraction μ.

But if we want to investigate for what wavelengths a given system of crystal planes gives rise to selective reflection, we must make use of (29), in which μ does enter. Thus the selective reflection occurs for wavelengths *different* from those for which it occurs in X rays; but for each of these the reflected beam has the same direction that it would have if we dealt with X rays of the same wavelength.

As we know, a great number of crystal planes exist in a crystal. In practice, however, only those in which the atoms lie sufficiently close together are effective in diffraction. To each of these corresponds a value for the right-hand side of equation (28) that is a constant characteristic of the system of crystal planes.

Therefore if we direct the rays normally upon the crystal in the manner described, we shall in general not have any diffracted beam, except for the case in which λ' has a value such that, when multiplied by an integer n, it gives a value coinciding with the characteristic constant of an actually effective crystal plane. Hence the experiment must be performed in the following manner. Fixing a value for the accelerating potential of the electrons, and thus for their velocity, and for λ', we investigate the distribution of the reflected electrons by bringing the collector C into various positions. In general we shall find a continuous distribution (a phenomenon analogous with that of optical scattering, which is always superimposed upon diffraction). Then we vary the potential, and hence λ', and again determine the distribution, and so forth until a potential is found for which a maximum occurs in a certain direction, or else until diffraction is superimposed upon the scattering. Equation (31) then permits us to find λ, and we have the wavelength corresponding to a given potential and hence to a given velocity.

Knowing λ, we can calculate the index of refraction μ corresponding to that wavelength by means of equation (29).

Figure 13 represents the results of such a series of experiments. Each polar diagram is obtained by laying off in each direction a distance proportional to the number of electrons scattered in that direction, and the various diagrams correspond to different velocities of the incident electrons. As can be seen, for a velocity of 54 volts there appears a well-defined diffracted beam in the direction $\theta = 50°$.

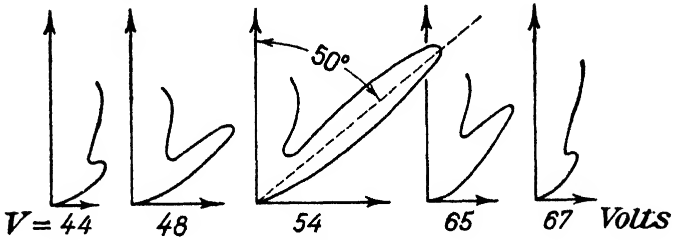

Fig. 13

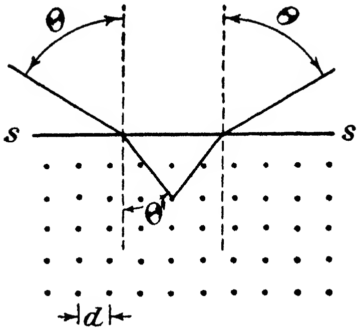

Fig. 14

31. Experiments performed by the Bragg method.

Davisson and Germer have also performed some measurements by a method analogous to the one used by Bragg with X rays.

The electrons are made to hit the crystal, not normally but at an angle of incidence θ (Fig. 14). The direction of propagation, in the interior of the crystal, makes an angle θ' with the normal to the surface, given by

$$\sin \theta' = \frac{\sin \theta}{\mu}. \tag{32}$$

The waves reflected from various crystal planes parallel to the surface ss as a rule interfere destructively, except for the case where Bragg's relation is satisfied:

$$n\lambda' = 2d \cos \theta', \tag{33}$$

where d is the distance between the crystal planes, measured perpendicular to the surface, and $\lambda' = \lambda/\mu$.

Solving for $\cos\theta'$ from (32) and substituting it into (33), we obtain

$$n\lambda' = \frac{2d}{\mu}\sqrt{\mu^2 - \sin^2\theta},$$

and hence

$$n\lambda = 2d\sqrt{\mu^2 - \sin^2\theta}. \tag{34}$$

In Bragg's experiments with X rays the crystal was made to oscillate so as to vary θ over a certain range, and a regular reflection was produced every time θ passed through a value satisfying (34). However, in the experiments with electrons it is more convenient to keep the angle of incidence θ fixed and to vary λ continuously by varying the accelerating potential V of the electrons.

Davisson and Germer always used a nickel crystal cut along a (111) plane. Upon drawing the polar plot of the electron distribution in various directions, for different values of V, they found that there generally exists an almost uniform distribution, except for some values of V for which quite a pronounced maximum is found in one direction, corresponding exactly to the law of optical reflection.

If the index of refraction of the crystal for the electron waves were 1, equation (34) would be reduced to the Bragg relation. That is, regular reflection would be obtained for those values of λ for which $n\lambda = 2d\cos\theta$, or else

$$\frac{1}{\lambda} = n\frac{1}{2d\cos\theta} = \text{const.} \times n. \tag{35}$$

Figure 15 shows a diagram of the intensities obtained in the direction corresponding to regular reflection (with an angle of incidence of 10°) as a function of $1/\lambda$, λ being calculated from the potential V by means of (26). The arrows indicate the points where the maxima should fall if (35) were valid. The influence of the index of refraction can well be seen from the figure: though (35) is sufficiently well verified for short wavelengths, because in that case the index of refraction is close to unity, the maxima for large wavelengths are considerably displaced. This displacement may be used to calculate μ as a function of λ, or else to establish experimentally for the de Broglie waves a law analogous to the law of optical dispersion.

Later experimenters in electron diffraction, using methods similar to those of Laue or Bragg and employing crystals of different substances, obtained results in essential agreement with those described above.

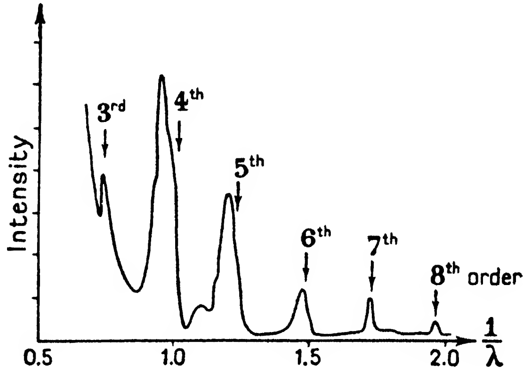

Fig. 15

32. Experiments performed by the Debye-Scherrer method. The method of Debye and Scherrer for the analysis of X rays and the study of the structure of crystals consists, as is well known, in letting a thin pencil of X rays fall on a crystalline material in powder form and in observing the diffracted rays. Since in the powder small crystals are found oriented in all directions, there will always be some crystals for which relation (35) is satisfied and which give rise to diffracted rays forming a cone whose axis lies along the direction of the incident beam. Consequently on a photographic plate placed beyond the tube containing the powder, there will be concentric rings from whose radii the wavelengths present in the beam may be deduced.

This method has been applied to electrons by G. P. Thomson, Rupp, and others. From many points of view, it has proved to be more advantageous than that of Davisson and Germer. By means of the Debye-Scherrer method, Thomson has been able to extend the wavelength measurements to electrons considerably faster than those of Davisson and Germer, attaining potentials up to 20,000 volts. Rupp performed the same experiments with slow electrons (several hundred volts), detecting them either photographically or by the electrometer method.

33. Velocity dependence of the wavelength. Theoretical considerations suggest, as we shall see in §26 of Part II, that the wavelength λ should depend upon the velocity v of the electrons according to a law of the type

$$\lambda = \frac{a}{mv}, \tag{36}$$

where a is a constant. Now the velocity v is related to the accelerating potential V by the kinetic energy relation, which for velocities small compared with c is

$$eV = \tfrac{1}{2}\, mv^2. \tag{37}$$

Hence we see that (36) is equivalent to

$$\lambda = \frac{a}{\sqrt{2em}} \frac{1}{\sqrt{V}}, \tag{38}$$

or else, calling λ_A the wavelength in angstroms and V_v the potential in volts,

$$\lambda_A = 10^8 \sqrt{\frac{300}{2em}} \frac{a}{\sqrt{V_v}}. \tag{38'}$$

The validity of the formula[10]

$$\lambda_A = \frac{12.25}{\sqrt{V_v}}. \tag{39}$$

has been confirmed by the measurements of the wavelengths λ_A (in angstroms) as a function of the accelerating voltage V_v (in volts) performed by the various methods described. The precision to which this law is verified is shown in the diagram of Fig. 16, in which the values measured for λ are plotted as ordinates corresponding to values of $1/\sqrt{V_v}$ along the abscissa. The straight line is described by equation (39), and we can see that the experimental points lie quite close to it (within experimental errors).

By identifying (38′) with the experimental relation (39), we obtain for the constant a the value

$$a = \sqrt{\frac{2em}{300}}\, 12.25 \times 10^{-8},$$

---

[10] See, for instance, Nos. 30 and 35 of the Bibliography.

and upon introducing the values of e and m, we find that $a = 6.614 \times 10^{-27}$. Thus we are justified in identifying a with Planck's constant h ($= 6.6237 \times 10^{-27}$ erg-sec), after which (36) becomes the already mentioned relation (26) between λ and v, which we can therefore consider to be based upon solid experimental foundations.[11]

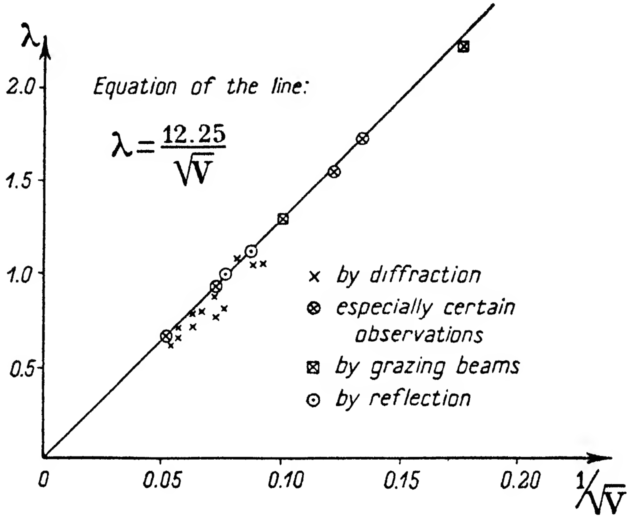

Fig. 16

It should be added that the preceding considerations are valid if the velocity v is small compared with c. Otherwise the kinetic energy law (37) becomes, according to the theory of relativity,

$$eV = m_0c^2\left(\frac{1}{\sqrt{1-(v^2/c^2)}} - 1\right), \tag{40}$$

and the mass m becomes a function of v according to the law

$$m = \frac{m_0}{\sqrt{1-(v^2/c^2)}}. \tag{41}$$

Nevertheless (26) still holds, if the expression (41) is inserted for m. Therefore the dependence of λ upon v and hence upon V becomes

---

[11] Independently of these, de Broglie had given a theoretical demonstration of (26) starting from first principles, based upon relativistic considerations (see No. 5 of the Bibliography).

somewhat more complicated.[12] The validity of (26), even for high velocities, has been verified by the experiments performed by Ponte with fast electrons (up to 17,250 volts) using the Debye-Scherrer method; he pushed the precision in the measurement of λ up to 3 parts per 1000 (the relativistic correction amounts to about 1 per cent for 10,000-volt electrons). More recently, Tappert[13] has extended the verification of the relativistic formula for λ up to 64,000-volt electrons, with an accuracy of 1 part per 1000.

34. Logical aspect of atomic mechanics. The operational definition of physical concepts. The electron diffraction experiments described in the preceding sections have clearly demonstrated that not only electromagnetic radiation but also so-called corpuscular rays, such as cathode rays, exhibit a double character; that is, they conform to a wave model in certain phenomena, to a particle model in others. Thus an even stricter parallelism is established between radiation and matter, and the problem of the connections between photons and electromagnetic waves (see §13) is transferred without change to the problem of the connections between electrons and de Broglie waves. In the case of radiation, phenomena of a wave nature were discovered first, then those of the corpuscular type; in the case of the electrons the reverse has taken place. It is only for this historical reason that we have become accustomed to conceive of radiation mainly in the wave aspect and of electrons in corpuscular form. The usual language still somewhat favors this tendency, since we speak more of light waves than of photons, more of electrons than of de Broglie waves. But in order to be perfectly objective it is necessary, in the case of electromagnetic radiation as well as in the case of cathode rays, to attribute equal importance to the properties of the wave type and to those of the particle type, and hence to convince ourselves that neither the wave nor the particle picture, as usually understood (that is, in analogy with the waves and particles visible to the eye), completely represents the nature of these physical entities.

---

[12] Instead, the dependence of λ upon *momentum* p is in every case the simple inverse proportionality; that is,

$$\lambda = \frac{h}{p},$$

since $p = mv$ even in relativistic mechanics.

[13] *Phys. Rev.* **54**, 1085 (1938).

For some time this fact presented an insurmountable difficulty, since it seemed that we had to deal with entities gifted with contradictory properties. The solution of this apparent contradiction was attained only through a process of logical clarification of the fundamentals of physics, which in turn came with the formal development of quantum mechanics. This development has been guided above all by the following logical fundamental criterion, which dominates, more or less explicitly, all the works of revision of physical principles; these works may be considered to have been initiated by the theory of relativity, and have been continued in a wider and more profound way in the new atomic mechanics. The objective of physics is to perform observations and experiments, and to coordinate their results in as simple a scheme as possible. Hence its immediate objectives are the observational data. But it is permissible to introduce other concepts for the interpretation of these data provided that such concepts be definable and detectable by means of experiments which, if not practicable, are at least *conceptually possible*, that is, not forbidden by any physical or logical precept. Quantities or relations that cannot be detected by means of such ideal but conceptually possible experiments may not be introduced into the reasoning without running the risk of being in error. Thus every concept of physics must be susceptible of an *operational definition;* that is, it must be definable by means of a series of conceptually possible physical operations. As one example, simultaneity is defined by indicating a method by which to decide whether two events are simultaneous or not. As another example, the electron is defined by indicating a way of detecting it; the coordinates of an electron, by indicating a procedure to measure them; and so forth. The only physical questions that have a meaning are those which ask for the result of one or more experiments which, at least conceptually, could be performed.

It is now appropriate to clear up the concept of conceptually possible experiments. These are not only experiments which may effectively be realized but also the limiting cases of actually possible experiments, provided that what prevents the limit from being reached is not a general law but a mass of practical difficulties. For example, perfectly reversible cycles have been used in thermodynamic reasoning for some time. This practice is permissible, since although such cycles may not be obtained in practice, there is still no theoretical reason why they should be impossible; and the more the techniques are perfected, the more closely these conditions are approached. Of course, the discovery of new general laws may modify certain of these opinions in the future by showing that an operation which at first was thought to be conceptually possible is not, or vice versa. Therefore, in speaking of conceptually possible operations, we imply "according to the physical laws now known." Making use only of such operations for the definition of concepts, one is at least certain not to introduce unwanted contradictions with known physical laws, and hence not to construct a logically incoherent structure, as may happen—and has happened many times—if concepts are introduced without being concerned with giving their operational definition. A famous example is the concept of *absolute* time, used in prerelativistic physics without a realization that its operational definition requires the instantaneous transmission of signals which, within the realm of hitherto known physical phenomena, is conceptually impossible. The point of departure of the relativity theory has indeed been the analysis of the concept of absolute time, and its operational definition. The same critical methods, when applied to atomic phenomena, lead to the uncertainty principle and to the new quantum mechanics.

The rigorous application of this criterion leads to more surprising results than could be imagined at first sight. The greatest benefit derived from it has been the clearing of the field of many questions that, upon an accurate analysis made with operational criteria, have proved to be devoid of physical meaning. Furthermore, it has enabled us to recognize that the paradoxical contradiction between the corpuscular and wave aspects of radiation, and the analogous contradiction more recently found in connection with matter, do not contain any actual logical contradiction. These apparent contradictions arise solely from our tendency to extend to the atomic domain those geometrical, kinematic, and causal representations which have been formed in our minds by means of the continual, unconscious experience accumulated through the observation of the macroscopic world, which we call "intuitions." We are thus led to introduce, into the atomic domain, concepts that are not capable of operational definition; hence the above-mentioned apparent contradictions. In other words, when speaking of particles (electrons or photons), we are led to imagine them as extraordinarily reduced material balls, and hence we are led to attribute to such entities all the geometrical and kinematic properties that ordinary bodies possess (such as continuity of trajectory), without realizing that in this manner we are postulating for the particles properties which are by no means experimentally proved, or even logically necessary. It is therefore not surprising that these properties are in contrast with those determined from experience. But if instead we associate with the concept of an electron or photon only properties that may be determined experimentally, we can verify that these properties, as is most natural, are not in logical contradiction with each other.

Hence there is no reason whatever why there should exist a geometrical or mechanical *model* accessible to our intuition and capable of completely interpreting *all the phenomena* of the microcosm. This is not to deny that models (such as the one of Rutherford-Bohr-Sommerfeld) may be in many cases of great use as a heuristic means as well as a concise means of expression, often furnishing exact, or at least approximate, interpretations of large classes of phenomena. However, they should not be taken literally but should be considered in the same light as hydrodynamic analogies, which aid so substantially in the understanding of electric current phenomena.

Furthermore, quantum mechanics also furnishes the explanation of the success of some of the geometrico-mechanical models in the interpretation of atomic phenomena. In fact, the laws of quantum mechanics retain some of the principles of ordinary mechanics, such as the principle of the conservation of energy and momentum. Besides, there is a continuous transition from the laws of micromechanics to those of macromechanics; the latter may, in many cases, serve as an approximate (or even exact) expression of the former.

35. The operator method and the Dirac theory of the electron. P. A. M. Dirac and P. Jordan have made a great contribution to the clarification of the logical aspect of atomic mechanics by working out the entire subject once more from an original point of view, based above all upon the physical concept of *state* and the mathematical concept of *operator*. They were able in this way to construct a unified and very general method of treatment for atomic problems (*transformation theory*), of which the matrix method and the method of wave mechanics are but particular cases.

Another remarkable result obtained by Dirac was the ability to show that, in the scheme of quantum mechanics, the mechanical angular momentum and the magnetic moment of the electron are necessary consequences of its existence and of the relativity principle, so that the so-called "electron-spin hypothesis" ceases to be a postulate in itself, since it is deduced as a consequence of the fundamental equations. This theory will be developed in Chapter 14.

Part II

WAVE MECHANICS OF A PARTICLE

CHAPTER 5

Mathematical Introduction

1. General remarks on linear homogeneous differential equations of the second order. In wave mechanics, differential equations (involving ordinary derivatives) which are linear, homogeneous and of the second order are of great importance; these are equations of the type

$$A(x)y'' + B(x)y' + C(x)y = 0, \tag{1}$$

where the coefficients A, B, and C are analytic functions of x, and are assumed to be real for real values of x. Since A is not supposed to be identically equal to zero, we can always write the equation in the form

$$y'' + P(x)y' + Q(x)y = 0, \tag{1'}$$

where we have put

$$P = \frac{B}{A}, \qquad Q = \frac{C}{A}.$$

First of all, we realize that any solution $y(x)$ is certainly regular (or else may be developed in a series of integral, positive powers of x) for all values of x for which the coefficients P and Q are regular. Only at points where at least one of these coefficients has a singularity may a singularity for y exist. Such points, called *singular points*, or *singularities* of the equation, are of considerable importance in the study of its properties. For the moment, however, we shall exclude them from our considerations, reserving the right to speak of them later on.

A well-known fundamental property of these equations is the following: once two particular solutions $y_1(x)$ and $y_2(x)$ which are independent (that is, whose ratio is not a constant) have been found, the general solution is found by forming a linear combination with them, using two arbitrary constants c_1 and c_2; this combination is

$$y(x) = c_1y_1(x) + c_2y_2(x). \tag{2}$$

We then say that y_1 and y_2 have been taken as *fundamental* *integrals.* Of course there is a large degree of arbitrariness in their choice, since we may replace them by any two of their linear combinations, provided that they are independent.

In actual problems, we must deal with the two constants c_1 and c_2 in such a way that y will satisfy two other conditions imposed by the problem: for example, that at a given point $x = a$, y and its derivative y' take on certain determined values (we see immediately that this is always possible and that y is thereby uniquely determined); or that y take on two given values at two given points a and b, or else that two given relations are to hold between the values of y and y' at points a and b. In many cases (but not always) such conditions in conjunction with the equation determine the function y uniquely. In general, the two points a and b lie on the real axis and are at the extremes of the interval within which the function $y(x)$ is of interest; therefore we speak of *boundary conditions.* We are mainly interested in the two following types of boundary conditions:

(α) y is to vanish at both extremes:

$$y(a) = 0, \qquad y(b) = 0; \tag{3}$$

(β) y is to assume the same values at the two extremes, and similarly y':

$$y(a) = y(b), \qquad y'(a) = y'(b). \tag{4}$$

Both of these types are contained in the category of "homogeneous conditions," to which many of the following considerations may be extended. They have the property that if $y(x)$ is a solution satisfying them, then $cy(x)$, where c is an arbitrary constant, is also a solution satisfying the same conditions.

We find that conditions (α) and (β) are satisfied, together with the differential equation, if we take $y = 0$ over the whole interval, or else if we make both constants of (2) equal to zero. This solution is evidently of no interest, so that our efforts will be directed toward finding nonzero solutions; that is, the case in which the boundary conditions in (α) or (β) *do not* uniquely determine a solution of the given equation is of interest. We shall see that this case is exceptional.

Let us consider, for instance, the case of conditions (α). Using expression (2) for the general solution, we are to find two values, not both zero, for c_1 and c_2 such that

$$\left.\begin{aligned} c_1y_1(a) + c_2y_2(a) &= 0 \\ c_1y_1(b) + c_2y_2(b) &= 0. \end{aligned}\right\} \tag{5}$$

It is known from algebra that this system of homogeneous equations has nonzero solutions only if

$$\begin{vmatrix} y_1(a) & y_2(a) \\ y_1(b) & y_2(b) \end{vmatrix} = 0. \tag{6}$$

This equation relates the values of the two fundamental integrals y_1 and y_2 at the points a and b; of course we obtain an equation of the same form if the solutions y_1 and y_2 are replaced by two other fundamental integrals.

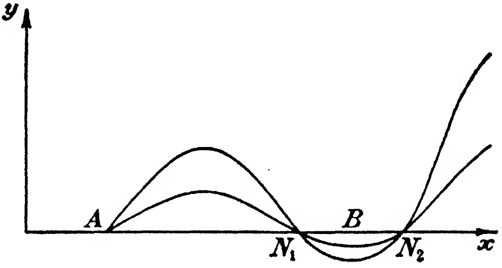

Fig. 17

If we can explicitly express $y_1(x)$ and $y_2(x)$ as functions of the coefficients of the equation, (6) immediately takes the form of a condition which these coefficients must satisfy for the equation to yield a nonzero solution for the given boundary conditions (and hence an infinite number of nonzero solutions, obtainable from the latter upon multiplying by a constant factor). Although, in general, no such explicit expression for y_1 and y_2 exists, (6) always expresses, in implicit form, a condition (of a functional type) for the coefficients in question. By similar reasoning for the case of conditions (β), we find, instead of (6), the condition

$$\begin{vmatrix} y_1(a) - y_1(b) & y_2(a) - y_2(b) \\ y_1'(a) - y_1'(b) & y_2'(a) - y_2'(b) \end{vmatrix} = 0. \tag{7}$$

In the case of conditions (α), condition (6) may be expressed in an intuitive form in the following manner. Each solution of the given equation, in the real domain, can be represented graphically by means of a curve in the xy-plane (see Fig. 17). Specifying the value at one point and the tangent there is enough to specify the curve completely. Now let us consider the family of all solution curves originating from the point A $(x = a)$ in various directions. It is easy to show that if one of them is given, any other may be obtained from it by means of a dilation or contraction along the y-axis. For example, let $y = f(x)$ be the curve starting at A with a slope 1 $[f'(a) = 1]$; then $y = \lambda f(x)$ (where λ is a constant) will be another solution and will be represented by the curve starting from A with a slope λ. By giving to λ a convenient value we can thus obtain any one of the solution curves originating at A. It follows that all the curves starting from A cut the x-axis at the same points N_1, N_2, . . . (*nodes*) corresponding to the roots of the equation $f(x) = 0$. In other words, by varying the slope of the tangent at A we do not displace the nodes, whose position, since A is fixed, is determined solely by the nature of the equation—that is, by its coefficients. If we try in this way to find a solution satisfying conditions (α)—that is, such that its curve passes, in addition to A, through another predetermined point B $(x = b)$—we find that *in general* such a solution does not exist. It exists only if one of the nodes N_1, N_2, . . . falls at B; and in that case, of course, the infinite number of solution curves issuing from A will satisfy the required condition. Now the condition that one of the nodes coincide with B, or that $f(b) = 0$, is equivalent to the condition found above in the form of equation (6).

Note. In the preceding argument y was supposed to be real, but it may be observed that every complex solution of (1) satisfying (α) or (β) at the extremes is composed of a real and an imaginary part, which satisfy the equation and the boundary conditions separately; therefore the reasoning can be applied to each of these parts.

2. Equations containing a parameter. Eigenvalues and eigenfunctions. Many cases of interest in wave mechanics and, in general, in the theory of oscillations of any kind, contain a *parameter* (that is, an undetermined constant) λ in the coefficients of the differential equation (1) or (1′). The most interesting case is the one in which λ occurs linearly in the coefficient C, so that the equation may be written in the form

$$A(x)y'' + B(x)y' + [\lambda\alpha(x) + \beta(x)]y = 0. \tag{8}$$

The fundamental integrals y_1 and y_2 will then also contain the parameter λ, and hence the left-hand member of (6) or of (7), which we shall designate by $\Delta(\lambda)$, will also contain it. Thus there appears the possibility of determining λ in such a way as to satisfy (6) or (7), and hence there may exist solutions of the proposed differential equation satisfying the desired boundary conditions (α) or (β). For this purpose, λ must be a root of the (generally transcendental) equation:

$$\Delta(\lambda) = 0. \tag{9}$$

Such values of λ are called the *eigenvalues* of the differential equation, relative to the interval (a, b) and to the boundary conditions (α) or (β). Once we have found an eigenvalue λ_i, relations (5) or similar ones allow us to find c_1 and c_2 in such a way that y satisfies the desired boundary conditions. [We observe that once (6) is satisfied, equations (5) are a consequence one of the other, so that they determine only the ratio c_2/c_1, and we may choose c_1 at will.] Such an integral of the differential equation is called an *eigenfunction* belonging to the eigenvalue λ_i.

It should be noted, however, that except for special cases, we cannot, in practice, determine the eigenvalues and eigenfunctions in this manner, since in order to put down equation (9), it is necessary to know the general solution of the differential equation. The above considerations serve only to clarify the concept of eigenvalues and eigenfunctions and to make their actual existence plausible. The problem of their determination is a fundamental one in wave theory. No general methods for its solution are known, only artifices that are successful in many cases. Some of these artifices will be applied in what follows.

3. Self-adjoint form of an equation. A differential equation of the type (1) is said to be in the *self-adjoint* form if between the coefficients $A(x)$ and $B(x)$ there exists the relation $B = A'$, so that the first two terms together form the derivative of Ay', and the equation may be written

$$\frac{d}{dx}(Ay') + Cy = 0. \tag{10}$$

It should be noted that any equation of the type (1) can be put into self-adjoint form. In fact, if we multiply (1) by the nonzero factor

$$\frac{1}{A}\, e^{\int (B/A)dx},$$

the three coefficients A, B, and C become, respectively,

$$P = e^{\int (B/A)dx}, \qquad Q = \frac{B}{A}\, e^{\int (B/A)dx}, \qquad R = \frac{C}{A}\, e^{\int (B/A)dx}, \tag{11}$$

and we can verify immediately that $Q = P'$, so that the equation can be written

$$\frac{d}{dx}\,(Py') + Ry = 0. \tag{12}$$

If then C contains a parameter λ linearly, as in (8), so will R, and equation (12) will have the form

$$(Py') + [\lambda\rho(x) + q(x)]y = 0. \tag{13}$$

Now the case which occurs most frequently is the one in which ρ is a constant, which may be taken equal to 1, so that the equation, reduced to its self-adjoint form, becomes

$$\frac{d}{dx}\,(Py') + [\lambda + q(x)]y = 0, \tag{14}$$

where P is always positive, as a result of (11). In what is to follow we shall refer in general to equations of this type, except when otherwise stated. Nevertheless, it should be remarked that the presence of the coefficient ρ would not introduce substantial modifications (at least if $\rho > 0$ over the whole interval considered) but only a slight complication of the formulas.

It can be shown[1] that an equation of type (14), with boundary conditions (α), always possesses an infinite number of real eigenvalues, and that this infinity is *discrete,* that is, denumerable, as long as the interval (a, b) is finite, as has been assumed here. Furthermore, if P is positive over the whole interval, the succession of eigenvalues is limited from below but has infinity for upper limit. To each of these eigenvalues there correspond one or, at most, two linearly independent eigenfunctions (never more than two, since any other solution can be expressed as a linear combination of these two, as is well known).

---

[1] Cf. No. 25 of the Bibliography.

4. Normalization of the eigenfunctions. Since any eigenfunction may be multiplied by an arbitrary constant without ceasing to satisfy the required conditions, there exists for each of them an infinity of other eigenfunctions, not independent of each other, which it is not of interest to consider as distinct solutions. To remove this indeterminacy, it is customary to add another condition, which is called *normalization* condition. This is

$$\int_a^b |y|^2\,dx = 1,$$

or else (if we designate, as we shall often do, the complex conjugate by an asterisk)

$$\int_a^b yy^*\,dx = 1. \tag{15}$$

This condition can always be satisfied, since, given a $Y(x)$ which is an eigenfunction not satisfying it, we have only to divide the eigenfunction by the nonzero constant

$$N = \sqrt{\int_a^b YY^*\,dx}$$

(*normalization factor*), since the eigenfunction $y = \frac{1}{N}Y$ obtained in this manner satisfies (15).

It should be noted that (15) is also satisfied if we multiply y by a complex constant of modulus 1 (that is, by a factor of the form $e^{i\theta}$). The normalization therefore does not completely determine the eigenfunction. Nevertheless, for most eigenfunctions we ignore this arbitrariness, which has no influence at all upon the modulus of the eigenfunction (see footnote on page 171).

5. Orthogonality of the eigenfunctions. We shall now show a fundamental property of the eigenfunctions.

Let us consider two eigenfunctions y_n and y_m of (14), obeying conditions (α) and belonging to two *distinct* eigenvalues λ_n and λ_m. They will identically satisfy the two relations

$$\frac{d}{dx}(Py_n') + (\lambda_n + q)y_n = 0, \tag{16}$$

$$\frac{d}{dx}(Py_m') + (\lambda_m + q)y_m = 0. \tag{16'}$$

The second relation will also be satisfied by y_m^*, the complex conjugate of y_m (the coefficients being real); that is, we obtain

$$\frac{d}{dx}(Py_m^{*\prime}) + (\lambda_m + q)y_m^* = 0.$$

Multiplying through by y_n, and (16) by y_m^*, and subtracting, we get

$$y_m^* \frac{d}{dx}(Py_n') - y_n \frac{d}{dx}(Py_m^{*\prime}) + (\lambda_n - \lambda_m)y_n y_m^* = 0.$$

We can immediately verify that the two first terms together are the exact derivative of $P(y_m^* y_n' - y_n y_m^{*\prime})$, so that, by multiplying the entire equation by dx and integrating between a and b, we have

$$[P(y_m^* y_n' - y_n y_m^{*\prime})]_a^b + (\lambda_n - \lambda_m) \int_a^b y_n y_m^* \, dx = 0.$$

Since at the limits both y_n and $y_m^{*\prime}$ vanish, the first part is zero. As long as we suppose that $\lambda_n \neq \lambda_m$, we are left with

$$\int_a^b y_n y_m^* \, dx = 0. \tag{17}$$

Such a relation, of the integral type, between the two eigenfunctions y_n and y_m is called (for reasons which will be explained in Chapter 10) an *orthogonality* relation. Thus we can say that *two eigenfunctions of* (14) *which vanish at the end points and which belong to different eigenvalues are orthogonal.*

The same property evidently still exists if the boundary conditions are of type (β), since the coefficient P takes on the same values at a and b.

We observe that (17) implies the conjugate relation

$$\int_a^b y_n^* y_m \, dx = 0. \tag{17'}$$

The orthogonality relations and normalization conditions can be expressed by a single formula by introducing the often used and very convenient symbol δ_{nm} with the meaning

$$\delta_{nm} = \begin{cases} 1 \text{ for } m = n \\ 0 \text{ for } m \neq n. \end{cases}$$

Then (15) and (17) are combined in the formula

$$\int_a^b y_n y_m^* \, dx = \delta_{nm}. \tag{18}$$

6. Degeneracy (multiple eigenvalues). As we have said, one or two linearly independent eigenfunctions may correspond to a given eigenvalue. If only one corresponds to it, all other solutions belonging to this eigenvalue are obtainable from this solution by multiplying it by a constant. If, however, the normalization condition is added, we can say that to the eigenvalue there corresponds a *single* normalized eigenfunction (apart from the constant of modulus 1, mentioned in §4). In this case the eigenvalue is said to be *simple*, since it is a simple root of equation $\Delta(\lambda) = 0$.

Let us now consider the case[2] in which two linearly independent eigenfunctions correspond to one eigenvalue. Here the eigenvalue is called *double;* and it will be said, in accordance with an expression that has become customary in atomic physics, that there is *degeneracy*. Let these two normalized eigenfunctions be $Y_1(x)$ and $Y_2(x)$ (we must remember that they will not in general be orthogonal, since the argument in §5 applies only to eigenfunctions belonging to two distinct eigenvalues). It is clear that from these eigenfunctions we may obtain an infinite number of other eigenfunctions corresponding to the same eigenvalue, by forming a linear combination with two arbitrary coefficients. We shall now prove that it is always possible, in an infinite number of ways, to construct two of these eigenfunctions which are *normalized* and *orthogonal.*

In fact, a first pair of orthogonal eigenfunctions can be built up of Y_1 itself and an appropriate linear combination $\overline{Y}_2 = \alpha Y_1 + \beta Y_2$. It suffices to choose the coefficients α and β such that

$$\int_a^b (\alpha Y_1 + \beta Y_2) Y_1^* \, dx = 0,$$

that is,

$$\alpha + \beta \int_a^b Y_2 Y_1^* \, dx = 0,$$

which fixes the ratio of α to β. One can then dispose of one of these to get $\overline{Y}_2$ to be normalized also. From the pair Y_1, $\overline{Y}_2$ there may then be obtained an infinite number of others by the formulas

$$y_1 = c_{11} Y_1 + c_{12} \overline{Y}_2,$$
$$y_2 = c_{21} Y_1 + c_{22} \overline{Y}_2;$$

and we find immediately that in order for y_1 and y_2 to be orthogonal and normalized, the coefficients must be subject to the restrictions

---

[2] It is immediately apparent that this case may occur with boundary conditions (β) but not with (α).

$$\left.\begin{aligned} c_{11}c_{11}^* + c_{12}c_{12}^* &= 1, \\ c_{21}c_{21}^* + c_{22}c_{22}^* &= 1, \\ c_{11}c_{21}^* + c_{12}c_{22}^* &= 0. \end{aligned}\right\} \tag{19}$$

In the case of real coefficients, these are simply the characteristic relations that exist, in analytic geometry, between the coefficients of the formulas for the rotation of the coordinate axes. These coefficients are known to be

$$\left.\begin{aligned} c_{11} &= \cos\theta, & c_{12} &= \sin\theta, \\ c_{21} &= -\sin\theta, & c_{22} &= \cos\theta. \end{aligned}\right\} \tag{20}$$

For this reason, a linear transformation of coefficients satisfying equations (19) is called an *orthogonal transformation.*

We may therefore say that to a double eigenvalue there correspond an infinite number of normalized pairs of eigenfunctions orthogonal to one another (besides being orthogonal to the other eigenfunctions), which pairs are obtained from one another by means of a general orthogonal transformation.

It is often convenient to consider a double eigenvalue [double root of the equation $\Delta(\lambda) = 0$] as two coinciding eigenvalues λ_1 and λ_2, and to make an eigenfunction that is normalized and orthogonal to the other correspond to each of them in the manner explained above. With this convention we can say that to every eigenvalue there corresponds an eigenfunction, all such eigenfunctions being orthogonal to each other.

7. Nodes of the eigenfunctions. In many questions it is of interest to examine those values of x for which the eigenfunction vanishes (*nodes* of the eigenfunction). We shall therefore state (without proof) a theorem concerning this matter (see Nos. 25 or 34 of the Bibliography).

Let us consider any solution of equation (14) which vanishes at A, and let it be graphically represented by one of the curves of Fig. 17. We have already mentioned (§1) that the position of the nodes depends only on the coefficients of the equation. If we now vary the parameter λ, which is contained in them, the nodes will be displaced, and it can be shown that when λ is being *increased continuously, the nodes are continuously displaced toward the left.* Every time one of these nodes coincides with the point B, the curve will satisfy the conditions (α) and hence will represent an eigenfunction, and the corresponding value of λ will be an eigenvalue, relative to conditions (α).

From this statement we can see that every time λ, upon increasing, passes through one of the eigenvalues, a new *node* enters the interval AB. If the (supposedly simple) eigenvalues are indicated by $\lambda_1, \lambda_2 \dots$ in increasing order, there will be no nodes in the interval AB as long as $\lambda < \lambda_1$; for $\lambda_1 < \lambda < \lambda_2$ there is one node, for $\lambda_2 < \lambda < \lambda_3$ there are two nodes, and so on. In particular, the eigenfunction y_n has $(n - 1)$ nodes in the interval AB, not counting the two at the end points; or, more clearly, *the nodes of the eigenfunction y_n divide the interval AB into n parts.*

8. The equation of harmonic motion. We shall illustrate what we have said by means of the following equation, well known in mechanics and called "equation of harmonic motion," with which we shall be concerned later on:

$$y'' + \lambda y = 0. \tag{21}$$

This equation evidently belongs to the type (14). In this case the procedure of §1 is actually applicable, since the general solution of (21) is known:

$$y_1 = c_1\, e^{i\sqrt{\lambda}\, x} + c_2\, e^{-i\sqrt{\lambda}\, x}. \tag{22}$$

The interval over which the integration is of interest is to be $(-l, l)$. Let us consider the two types of boundary conditions (α) and (β) separately.

Conditions (α): We must have[3]

$$y(-l) = y(l) = 0,$$

that is,

$$\begin{aligned} c_1\, e^{-i\sqrt{\lambda}\, l} + c_2\, e^{i\sqrt{\lambda}\, l} &= 0, \\ c_1\, e^{i\sqrt{\lambda}\, l} + c_2\, e^{-i\sqrt{\lambda}\, l} &= 0. \end{aligned} \tag{23}$$

In order that this system of linear homogeneous equations in c_1 and c_2 may have nonzero solutions, we must have

$$\begin{vmatrix} e^{-i\sqrt{\lambda}\, l} & e^{i\sqrt{\lambda}\, l} \\ e^{i\sqrt{\lambda}\, l} & e^{-i\sqrt{\lambda}\, l} \end{vmatrix} = 0.$$

Upon expanding, this equation leads to

---

[3] These conditions, for instance, come up in the problem of standing waves in a string with fixed endpoints. Another example will be seen in §38.

$$\sin 2\sqrt{\lambda}l = 0,$$

or else

$$2\sqrt{\lambda}l = n\pi \quad (n = 0, 1, 2, \ldots). \quad (24)$$

Therefore λ must have one of the values

$$\lambda_n = n^2\frac{\pi^2}{4l^2} \quad (24')$$

which are the desired eigenvalues.

λ having been determined, we get from either equation (23)

$$c_1 = c_n\, e^{i\sqrt{\lambda_n}\,l}, \qquad c_2 = -c_n\, e^{-i\sqrt{\lambda_n}\,l}, \quad (23')$$

where c_n is an arbitrary constant. Substituting into (22), we get

$$y_n = c_n 2i \sin\sqrt{\lambda_n}\,(x + l) = C_n \sin\frac{n\pi}{2l}(x + l),$$

where the factor $2i$ has been incorporated into the arbitrary constant c_n. Except for the irrelevant case in which $n = 0$ (that is, in which y vanishes identically), the normalization condition then yields,

$$\int_{-l}^{l} y_n y_n^*\, dx = lC_nC_n^* = 1,$$

whence

$$|C_n| = \frac{1}{\sqrt{l}}.$$

If we require y_n to be real, we must take for the eigenfunctions

$$y_n = \sqrt{\frac{1}{l}}\sin\frac{n\pi}{2l}(x + l). \quad (25)$$

We can then immediately verify the orthogonality theorem, since for $n \neq m$ we have

$$\int_{-l}^{l} \sin\frac{n\pi}{2l}(x + l)\sin\frac{m\pi}{2l}(x + l)\, dx = 0.$$

The existence of discrete eigenvalues in this problem may be explained intuitively in the following manner. The solutions of (21) are graphically represented by exponential curves if $\lambda < 0$, and by sinusoidal curves if $\lambda > 0$. The first have at most one node; the latter have an infinite number of nodes spaced a distance $\pi/\sqrt{\lambda}$ apart. If we require that two nodes are to fall at the two end points, we must discard the exponentials and equate the interval $2l$ to a multiple of the distance between two nodes; this yields (24). The nth eigenfunction (25) is represented by a sine curve that possesses, in addition to the two nodes at A and B, $(n - 1)$ intermediate nodes (see §7). The normalization condition then determines the amplitude of the sine function.

Conditions (β): We must then have

$$\begin{aligned} y(-l) &= y(l), \\ y'(-l) &= y'(l), \end{aligned}$$

that is,

$$\left.\begin{aligned} c_1\, e^{-i\sqrt{\lambda}l} + c_2\, e^{i\sqrt{\lambda}\,l} &= c_1\, e^{i\sqrt{\lambda}\,l} + c_2\, e^{-i\sqrt{\lambda}\,l} \\ c_1\, e^{-i\sqrt{\lambda}l} - c_2\, e^{i\sqrt{\lambda}\,l} &= c_1\, e^{i\sqrt{\lambda}\,l} - c_2\, e^{-i\sqrt{\lambda}\,l} \end{aligned}\right\} \tag{26}$$

These conditions may also be written

$$\left.\begin{aligned} (c_1 - c_2) \sin \sqrt{\lambda}l &= 0, \\ (c_1 + c_2) \sin \sqrt{\lambda}l &= 0. \end{aligned}\right\} \tag{27}$$

Hence, if c_1 and c_2 are not both zero,

$$\sin \sqrt{\lambda}l = 0,$$

from which

$$\sqrt{\lambda}l = n\pi \qquad (n = 0, 1, \ldots). \tag{28}$$

The eigenvalues are therefore

$$\lambda_n = n^2 \frac{\pi^2}{l^2}. \tag{28'}$$

If λ is chosen in this way, both boundary conditions will be satisfied, and hence c_1 and c_2 remain arbitrary. The condition of normalization yields only

$$c_1c_1^* + c_2c_2^* = \frac{1}{2l}.$$

We therefore have a case of degeneracy: the eigenvalues are double (except for the eigenvalue 0).

We may obtain a pair of independent, orthogonal, and normalized solutions (for $n \neq 0$) by taking in (22)

$$c_1 = \frac{1}{\sqrt{2l}},\ c_2 = 0, \qquad \text{or} \qquad c_1 = 0,\ c_2 = \frac{1}{\sqrt{2l}},$$

respectively, and then

$$y_n^{(1)} = \frac{1}{\sqrt{2l}} e^{in(\pi/l)x}, \qquad y_n^{(2)} = \frac{1}{\sqrt{2l}} e^{-in(\pi/l)x}. \tag{29}$$

Another solution is obtained by taking

$$c_1 = \frac{1}{2\sqrt{l}}, \; c_2 = \frac{1}{2\sqrt{l}}, \qquad \text{or} \qquad c_1 = \frac{1}{2i\sqrt{l}}, \; c_2 = -\frac{1}{2i\sqrt{l}},$$

respectively, and then

$$y_n^{(1)} = \frac{1}{\sqrt{l}} \cos \frac{n\pi x}{l}, \qquad y_n^{(2)} = \frac{1}{\sqrt{l}} \sin \frac{n\pi x}{l}. \tag{30}$$

(For the eigenvalue 0 we have instead the unique eigenfunction $y_0 = 1/\sqrt{2l}$). From any one of these pairs we can get an infinite number of others by orthogonal transformation. For example, from (30), by the transformation (20), we obtain the pair

$$y_n^{(1)} = \frac{1}{\sqrt{l}} \cos \left(\frac{n\pi x}{l} - \theta\right), \qquad y_n^{(2)} = \frac{1}{\sqrt{l}} \sin \left(\frac{n\pi x}{l} - \theta\right)$$

with θ arbitrary.

We might have foreseen these results intuitively by observing that conditions (β) force the sine curve to have the same ordinate and the same slope at A and B [and hence the interval AB must be a multiple of the wavelength $2\pi/\sqrt{\lambda}$, whence (28)], but its phase θ at A remains arbitrary.

9. Expansion of a function in a series of eigenfunctions. Fourier series.[4] Let us consider the infinite sequence of *orthonormal* (orthogonal and normalized) eigenfunctions $y_1, y_2 \ldots$, corresponding to the eigenvalues of equation (14), with boundary conditions (α) or (β) and with the convention (explained in §6) that any double eigenvalues are to be counted twice. We can prove that, given any function $f(x)$ satisfying the same boundary conditions and also certain qualitative conditions which are not very restrictive, it is possible to expand it over the interval (a, b) in a series of the form

$$f(x) = \sum_{n=1}^{\infty} f_n y_n(x), \tag{31}$$

---

[4] See No. 32a of the Bibliography, §§74–77.

which is uniformly convergent in the interval (a, b).

Assuming this theorem,[5] we recognize immediately that the coefficients f_n are given by

$$f_n = \int_a^b f(x) y_n^* (x)\, dx; \tag{32}$$

indeed, upon multiplying both sides of (31) by $y_r^*(x)\, dx$ (where r is any one of the indices 1, 2, . . .) and integrating between a and b, all terms in the right-hand member will vanish by virtue of the orthogonality of the y, except the one for which $n = r$, which reduces to f_r. Upon changing the index from r to n, we obtain (32).

This procedure is recognizable as merely a generalization of the well-known procedure used to determine the coefficients of the *Fourier expansion;* the latter is just a particular case of expansion in a series of orthogonal functions, namely, of eigenfunctions of (21) with boundary conditions (β). In fact, taking these eigenfunctions in the form (30) and incorporating the constant $1/\sqrt{l}$ into the coefficients f_n (which we shall call a_n and b_n), we can write the development (31) as

$$f(x) = \sum_{n=0}^{\infty} \left(a_n \cos \frac{n\pi x}{l} + b_n \sin \frac{n\pi x}{l} \right) \tag{33}$$

---

[5] See No. 25 or 34 of the Bibliography. More generally the following theorem holds. If the function f is such that $\int_a^b |f|^2\, dx$ exists, the series (31) is convergent *at least in the mean,* so that

$$\lim_{N\to\infty} \int_a^b \left| f - \sum_{n=1}^{N} f_n y_n \right|^2 dx = 0. \tag{31*}$$

Expanding the square occurring in this equation and using (32) and the orthogonality property, we find the important formula of Parseval:

$$\int_a^b |f|^2\, dx = \sum_{n=1}^{N} |f_n f_n^*|, \tag{31**}$$

which expresses the *completeness* of the system of eigenfunctions y_n. In fact, considering any system of orthogonal functions, even infinite, but not representing the *totality* of eigenfunctions of a differential equation, the sign $>$ rather than $=$ will hold in (31**).

where, according to (32),

$$\left.\begin{aligned} a_0 &= \frac{1}{2l}\int_{-l}^{l} f(x)\,dx \\ a_n &= \frac{1}{l}\int_{-l}^{l} f(x)\cos\frac{n\pi x}{l}\,dx, \\ b_n &= \frac{1}{l}\int_{-l}^{l} f(x)\sin\frac{n\pi x}{l}\,dx. \end{aligned}\right\} \tag{34}$$

This is the *Fourier expansion.* It represents the function $f(x)$ over the interval $(-l, l)$ even if the latter has points of discontinuity there (at which the series represents the arithmetic mean of the two limits from the right and from the left). In order for the expansion to hold, it is sufficient that the interval be divisible into segments over each of which f is continuous and monotonic.

The same expansion may be obtained in a more convenient form by using the eigenfunctions (29), which can be collected into the single formula

$$y_n = \frac{1}{\sqrt{2l}}\,e^{i(n\pi x/l)}, \tag{35}$$

provided that we make the convention that n may also take on negative values. Then, when we set $f_n/\sqrt{2l} = C_n$, (31) yields (*Fourier expansion in exponential form*)

$$f(x) = \sum_{n=-\infty}^{\infty} C_n\,e^{i(n\pi x/l)} \tag{36}$$

and (32) yields the expression for the coefficients

$$C_n = \frac{1}{2l}\int_{-l}^{l} f(x)\,e^{-i(n\pi x/l)}\,dx \tag{37}$$

If we calculate, by means of (36), the integral of ff^* extended over the interval $(-l, l)$, we readily find that

$$\int_{-l}^{l} |f|^2\,dx = 2l\sum_{n=-\infty}^{\infty} |C_n|^2 \tag{38}$$

which is simply the *completeness relation* for the eigenfunctions in question.

10. Case of an infinite interval. Continuous spectrum of eigenvalues. In certain problems of wave mechanics, equation (14) must be integrated, not over a finite interval (a, b) as we have assumed thus far, but over an interval infinite on one or on both sides—(a, ∞) or $(-\infty, \infty)$. It will be convenient to examine this case as a limit of that in which integration is over (a, b) by making one or both endpoints tend toward infinity. It will then be recognized that the problem exhibits new characteristics in the case of the infinite interval, of which the most important is the appearance of a *continuous spectrum* of eigenvalues. With a continuous spectrum there may exist for λ (in addition to an eventual discrete sequence of eigenvalues) some intervals within which any value of λ is an eigenvalue; and, of course, to these eigenvalues there corresponds a continuous set of eigenfunctions, which for convenience we shall designate by $y(x)$ [where the index λ varies continuously, so that, actually, y is a function of two variables and may also be indicated by $y\ (\lambda, x)$].

We shall only consider the case in which the interval is $(-\infty, +\infty)$, and we shall confine ourselves to some considerations of an intuitive nature by using the example of §8, case (β). If we start with one of the eigenvalues (28′) and the corresponding eigenfunctions (29), and let l tend toward ∞, we shall see that λ_n approaches the eigenvalue 0 and that the two eigenfunctions also tend toward zero for any x; therefore we obtain nothing of interest. Instead, it is necessary, upon letting l tend toward ∞, to consider eigenfunctions of higher and higher order; in this process n will also approach infinity, in such a manner[6] that the eigenvalue $\lambda_n = n^2(\pi^2/l^2)$ approaches a predetermined limit $\lambda (> 0)$. The two eigenfunctions then approach

$$y_\lambda{}^{(1)} = a_\lambda\, e^{i\sqrt{\lambda}x}, \qquad y_\lambda{}^{(2)} = a_\lambda\, e^{-i\sqrt{\lambda}x}, \tag{39}$$

---

[6] In order to understand this process intuitively, let us imagine that the eigenvalues (28′) are represented by points on a straight line (Fig. 18). It is clear from the formula that when l is increased, these points are displaced indefinitely toward the origin, becoming more and more closely spaced. Through a fixed point on the line, there successively pass eigenvalues of higher and higher order; that is, a given value of λ is to be identified with eigenvalues of higher and higher order. Therein lies the significance of letting n approach ∞ at the same time as l.

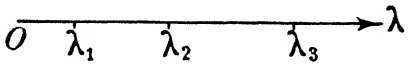

Fig. 18

where a_λ is a constant so far as x is concerned. In fact, if the two expressions are combined into one by setting $\omega = +\sqrt{\lambda}$ for the first and $\omega = -\sqrt{\lambda}$ for the second,

$$y_\omega = a_\omega\, e^{i\omega x}. \tag{39'}$$

We may consider this formula analogous to (35). The continuous variable ω, which may range from $-\infty$ to $+\infty$, takes the place of the discontinuous variable $n\pi/l$.

The extension of the orthogonality theorem and of normalization to these "eigenfunctions with continuous spectrum" requires some attention; in fact, it is obvious from (39′) that $y_\omega y_\omega^*$ does not approach zero for $x \to \pm\infty$, and hence that the normalization integral does not converge.

Instead, it is necessary to consider, not an eigenfunction belonging to a definite value of λ (we refer for the present only to $y^{(1)}$ or $y^{(2)}$), but the sum of all eigenfunctions belonging to an *infinitesimal interval* $(\lambda_0, \lambda_0 + \Delta\lambda)$, or the integral

$$\Delta y = \int_{\lambda_0}^{\lambda_0+\Delta\lambda} y_\lambda\, d\lambda \tag{40}$$

which is an infinitesimal of order $\Delta\lambda$, and is called *eigendifferential* (by the notation Δy, we imply a passage to the limit for $\Delta\lambda \to 0$). Setting $\lambda = \lambda_0 + \Delta\lambda/2 + \epsilon$ and neglecting the squares of ϵ and of $\Delta\lambda$, we easily find (writing λ instead of λ_0)[7]

$$\Delta y = a_\lambda\, e^{i\sqrt{\lambda+\frac{1}{2}\Delta\lambda}\,x}\,\frac{\sin\dfrac{x\,\Delta\lambda}{4\sqrt{\lambda}}}{\dfrac{x\,\Delta\lambda}{4\sqrt{\lambda}}}\,\Delta\lambda, \tag{41}$$

or else, setting $\sqrt{\lambda} = \omega$ as above,

$$\Delta y = a_\omega\, e^{i(\omega+\frac{1}{2}\Delta\omega)x}\,\frac{\sin\dfrac{x\,\Delta\omega}{2}}{\dfrac{x\,\Delta\omega}{2}}\,2\omega\,\Delta\omega, \tag{41'}$$

---

[7] Since the integral (40) is extended over an infinitesimal interval $\Delta\lambda$, we can consider that it will reduce to a single element and may be obtained by writing $\Delta\lambda$ instead of $d\lambda$ in the integrand. This result, however, holds only for finite values of x, as can be seen from expression (41), since even an infinitesimal variation of λ makes a difference in the value of the function, which becomes larger as x increases.

an expression which also extends to negative values of ω.

The function $\Delta y/\Delta\omega$ can be seen to approach 0 when $x \to \pm\infty$, a statement that can be interpreted by saying that the infinite number of superimposed sine functions cancel at infinity by mutual interference. For finite x, however, $\Delta y/\Delta\omega$ approaches y_ω as $\Delta\omega$ approaches zero.

In addition, the functions possess a property that is the natural extension of the orthogonality of two eigenfunctions belonging to different eigenvalues. In fact, if $\Delta_1 y$ and $\Delta_2 y$ are the Δy's belonging to two small intervals of eigenvalues $(\lambda_1, \lambda_1 + \Delta_1\lambda)$ and $(\lambda_2, \lambda_2 + \Delta_2\lambda)$ that do not even partially overlap, we have, as may be verified by a somewhat laborious integration,

$$\int_{-\infty}^{\infty} \Delta_1 y \, \Delta_2 y^* \, dx = 0. \tag{42}$$

If instead the two small intervals have a segment $\Delta\lambda$ in common, the integral proves to be

$$4\pi a_\lambda a_\lambda^* \sqrt{\lambda}\, \Delta\lambda,$$

where λ_1 and λ_2 are indifferently designated by λ, since they differ by an infinitesimal. This result suggests setting up as normalization condition

$$\frac{1}{\Delta\lambda}\int_{-\infty}^{\infty} \Delta_1 y \, \Delta_2 y^* \, dx = 1. \tag{43}$$

Taking into account the value of the integral shown above, we see that the normalization constant a_λ must then be chosen so that

$$a_\lambda a_\lambda^* = \frac{1}{4\pi\sqrt{\lambda}}.$$

We can then take as the normalized eigenfunction of (21) in the interval $(-\infty, +\infty)$

$$y_\lambda = \frac{1}{2\sqrt{\pi\sqrt{\lambda}}}\, e^{\pm i\sqrt{\lambda}\, x}, \tag{44}$$

or else

$$y_\omega = \frac{1}{2\sqrt{\pi\omega}}\, e^{i\omega x}. \tag{44'}$$

Generalizing what we have shown for this particular example, we can say that equation (14), for an infinite interval, can possess a continuous spectrum of eigenvalues. The eigenfunctions belonging to these eigenvalues do not, in general, approach zero at infinity, but their eigendifferentials Δy do. The boundary condition is precisely[8]

$$\lim_{x\to\pm\infty} \frac{1}{\Delta\lambda}\int_{\lambda}^{\lambda+\Delta\lambda} y_\lambda(x)\,d\lambda = 0 \tag{45}$$

and the orthogonality property is

$$\int_{-\infty}^{\infty} dx\left[\int_{\lambda_1}^{\lambda_1+\Delta_1\lambda} y_\lambda(x)\,d\lambda\cdot\int_{\lambda_2}^{\lambda_2+\Delta_2\lambda} y_\lambda{}^*(x)\,d\lambda\right] = 0 \tag{46}$$

for two intervals $\Delta_1\lambda$ and $\Delta_2\lambda$ that do not overlap. The normalization condition is

$$\frac{1}{\Delta\lambda}\int_{-\infty}^{\infty} dx\left[\int_{\lambda_1}^{\lambda_1+\Delta_1\lambda} y_\lambda(x)\,d\lambda\cdot\int_{\lambda_2}^{\lambda_2+\Delta_2\lambda} y_\lambda{}^*(x)\,d\lambda\right] = 1 \tag{46'}$$

for two small intervals which have a segment $\Delta\lambda$ in common, or which possibly coincide.

If, besides the continuous eigenvalues, there are also discrete eigenvalues λ_n, the following orthogonality property holds between the eigenfunctions of the continuous spectrum and those of the discrete spectrum:

$$\int_{-\infty}^{\infty} \frac{1}{\Delta\lambda}\, y_n(x)\,dx \int_{\lambda}^{\lambda+\Delta\lambda} y_\lambda{}^*(x)\,d\lambda = 0. \tag{47}$$

We can now extend the expansion of an arbitrary function in a series of eigenfunctions to the case of an infinite interval. However, in correspondence to continuous eigenvalues, we shall have an integral rather than a series. Referring, for greater generality, to the case in which both discrete and continuous eigenvalues exist, we have an expansion of the form[9]

$$f(x) = \sum_{n=1}^{\infty} f_n y_n(x) + \int f_\lambda y_\lambda(x)\,d\lambda, \tag{48}$$

---

[8] The limit $x \to \pm\infty$ refers to the case where the interval is infinite on both ends; otherwise we are to read only $+\infty$, or $-\infty$, and to substitute the ordinary condition $y = 0$ for the other endpoint. The integrations with respect to x are understood to be extended over the entire interval.

[9] Extending the concept of integral in the sense of Stieltjes, we could write this formula, like all analogous ones, with a single integral, making it also contain possible discrete terms (see, for example, page 123 of No. 14 of the Bibliography).

where f_λ is a function of the continuous variable λ, and the integration with respect to λ is understood to be carried out over the whole continuous spectrum of eigenvalues. The determination of the coefficients f_n is made as in §9, with theorem (47) used in addition to the orthogonality of the y_n, and one finds the same expression (32). The function f_λ (which appears under the integral sign in a manner analogous to the f_n under the summation sign) is similarly determined in the following manner. Let us first of all decompose the continuous spectrum into intervals $\Delta_1\lambda$, $\Delta_2\lambda$, . . . and let us break up the integral as follows:

$$f(x) = \sum_n f_n y_n(x) + \sum_i \int_{\lambda_i}^{\lambda_i+\Delta_i\lambda} f_\lambda y_\lambda(x)\, d\lambda.$$

We now multiply both sides of the equation by

$$\frac{dx}{\Delta_r\lambda} \int_{\lambda_r}^{\lambda_r+\Delta_r\lambda} y_\lambda^*(x)\, d\lambda$$

and integrate with respect to x (over the whole range of x). Then, letting the intervals $\Delta_r\lambda$ approach zero, we see that by virtue of (47), all terms of the first sum on the right-hand side vanish, and according to (46) all terms of the second sum vanish except the rth term, which reduces to f_{λ_r} according to (46′). Omitting the index r, since λ varies continuously, we have

$$f_\lambda = \lim_{\Delta\lambda\to 0} \frac{1}{\Delta\lambda} \int_{-\infty}^{\infty} f(x)\, dx \int_{\lambda}^{\lambda+\Delta\lambda} y_\lambda^*(x)\, d\lambda, \tag{49}$$

and if $f(x)$ becomes infinitesimal of sufficiently high order at infinity, we can interchange the limit with the integral sign and obtain

$$f_\lambda = \int_{-\infty}^{\infty} f(x) y_\lambda^*(x)\, dx. \tag{50}$$

The completeness relation (in the case of the continuous spectrum) is

$$\int_{-\infty}^{\infty} |f|^2\, dx = \int |f_\lambda|^2\, d\lambda. \tag{51}$$

11. Expansion in a Fourier integral. The most important application of the preceding considerations is the expansion in a *Fourier integral,* which is the natural extension of the Fourier series expansion to the case where it is desired to represent a given (nonperiodic) function $f(x)$ over an infinite interval. This function can be obtained by taking expressions (44) as eigenfunctions, and hence

$$f(x) = \int_0^\infty \frac{c_1(\lambda)}{2\sqrt{\pi}\sqrt{\lambda}} e^{i\sqrt{\lambda}\,x}\, d\lambda + \int_0^\infty \frac{c_2(\lambda)}{2\sqrt{\pi}\sqrt{\lambda}} e^{-i\sqrt{\lambda}\,x}\, d\lambda, \tag{52}$$

or else

$$f(x) = \int_{-\infty}^{\infty} C(\omega)\, e^{i\omega x}\, d\omega, \tag{53}$$

having set

$$\left.\begin{aligned} C(\omega) &= \sqrt{\frac{\omega}{\pi}}\, c_1(\omega^2) \qquad &(\text{for } \omega > 0), \\ \text{and} \qquad C(\omega) &= -\sqrt{\frac{-\omega}{\pi}}\, c_2(\omega^2) \qquad &(\text{for } \omega < 0). \end{aligned}\right\}$$

The coefficients c_1 and c_2 are given by (50), and we find

$$C(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(x)\, e^{-i\omega x}\, dx. \tag{54}$$

We observe that if $f(x)$ is an even function, that is, if $f(-x) = f(x)$, then $C(\omega)$ will also be even, and if $f(x)$ is odd, $C(\omega)$ will be odd. In the first case, (53) can also be written as

$$f(x) = 2\int_0^\infty C(\omega)\cos\omega x\, d\omega, \tag{53$'$}$$

with

$$C(\omega) = \frac{1}{\pi}\int_0^\infty f(x)\cos\omega x\, dx; \tag{54$'$}$$

and in the second case:

$$f(x) = 2i\int_0^\infty C(\omega)\sin\omega x\, d\omega, \tag{53$''$}$$

with

$$C(\omega) = \frac{1}{i\pi}\int_0^\infty f(x)\sin\omega x\, dx. \tag{54$''$}$$

Even in the general case, (53) and (54) can easily be put into trigonometric form.

12. Wave interpretation of the Fourier expansion. The expansion in a Fourier integral is susceptible of an expressive physical interpretation which may apply to wave phenomena of any kind, including elastic waves, light waves, sound waves, and others. We shall investigate this interpretation by using the terminology of electromagnetic optics, but the translation of the same considerations into the language corresponding to other types of wave phenomena is immediate.

If monochromatic radiation of wavelength λ travels (in plane waves) with a velocity V along the x-axis, any one of the components f of the electric or magnetic field is represented, as a function of t and of x, by the well-known *wave equation*

$$f(x,t) = A\,e^{i\frac{2\pi}{\lambda}(x-Vt)} = A\,e^{2\pi i(kx-\nu t)} \tag{55}$$

where we set
$$k = \frac{1}{\lambda} \qquad (\text{"wave number"})$$
and
$$\nu = \frac{V}{\lambda} \qquad (\text{"frequency"})$$

and where A is a generally complex constant whose absolute value represents the amplitude of the waves, whereas its argument represents their phase. We shall call A the *complex amplitude* [as usual the physical quantity f may be understood to be represented by the real part, or by the imaginary part, of expression (55)].

In any case, f is represented graphically by a sinusoidal curve of wavelength λ along the x-axis, which progresses in time in the positive direction of the x-axis (if $V > 0$) with the uniform velocity V. We shall say that f represents a monochromatic *wave train.*[10] We then note that a *reverse* wave train may be represented by the same expression, when k is given a negative value.

Heterochromatic radiation will be represented by a sum of terms of the type (55) (with different values of λ and hence different values of k and ν), each corresponding to a single monochromatic component or to a single *spectral line.* In general, the velocity V will also be different for the various radiation components; that is, there will be *dispersion.* We can select any one of these quantities λ, k, and ν to characterize the individual monochromatic components. We shall take the "wave number k" and shall suppose that the function $V(k)$, governing the dispersion of the medium, is known. The frequency will then be given by

$$\nu(k) = \frac{V(k)}{\lambda} = kV(k). \tag{56}$$

---

[10] By the word "train" we designate a sequence of waves unlimited in space (from $-\infty$ to $+\infty$) and hence in time. Instead, we shall use the word "group" when dealing with a limited number of waves.

If the radiation, when observed with a spectroscope, produces a continuous spectrum rather than a line spectrum, we must represent it by an integral rather than a sum, which will be

$$f(x, t) = \int_{-\infty}^{\infty} A(k)\, e^{2\pi i(kx-\nu t)}\, dk. \tag{57}$$

The quantity $|A|^2\, dk$ represents the luminous intensity corresponding to the interval dk of the continuous spectrum; hence the function $|A(k)|^2$ represents the intensity distribution in the spectrum. In order to represent, by a single formula, the superposition of radiation traveling in the forward direction and radiation traveling in the opposite direction, we have extended the integral in (57) from $-\infty$ to $+\infty$, with the convention that $A(-k)$ represents the amplitude and phase of the waves of wave number k traveling in the opposite direction.

At time zero, the distribution of f along the x-axis is given by

$$f(x, 0) = \int_{-\infty}^{\infty} A(k)\, e^{2\pi ikx}\, dk, \tag{58}$$

a formula which may be identified with (53), provided that we put $\omega = 2\pi k$, $C(\omega) = \frac{1}{2\pi} A\left(\frac{\omega}{2\pi}\right)$. With these substitutions, (54) yields the following expression for A:

$$A(k) = \int_{-\infty}^{\infty} f(x, 0)\, e^{-2\pi ikx}\, dx, \tag{59}$$

and the completeness relation (51) is written in this case as

$$\int_{-\infty}^{\infty} |f(x, t)|^2\, dx = \int_{-\infty}^{\infty} |A|^2\, dk. \tag{51$'$}$$

Thus, given the initial distribution of f, Fourier's theorem points out how to decompose it into a superposition of infinite monochromatic wave trains whose complex amplitudes are given by (59). These wave trains then travel, each with a different velocity $V(k)$, and their superposition, at time t, gives rise to a function $f(x, t)$ (in general represented by a curve of different shape from the initial one) expressed by (57).

In the case where $f(x, 0)$ is an even (or odd) function, formulas (53′), (54′), (53″), (54″) may be modified to yield for even f:

$$f(x, 0) = 2\int_0^\infty A(k) \cos 2\pi kx \, dk, \tag{58'}$$

$$A(k) = 2\int_0^\infty f(x, 0) \cos 2\pi kx \, dx; \tag{59'}$$

and for odd f:

$$f(x, 0) = 2i\int_0^\infty A(k) \sin 2\pi kx \, dk, \tag{58''}$$

$$A(k) = \frac{2}{i}\int_0^\infty f(x, 0) \sin 2\pi kx \, dx. \tag{59''}$$

Example: Group of waves of constant amplitude. We shall apply the preceding considerations to the case in which the initial distribution of f is the one represented by Fig. 19, that is,

$$\begin{cases} f(x, 0) = \sin 2\pi k_0 x & (\text{for } -l < x < l) \\ f(x, 0) = 0 & (\text{for } x < -l,\ x > l). \end{cases}$$

We say that we are dealing with a *wave group* of length $2l$. We can imagine it to be produced approximately by means of a source which is capable of emitting monochromatic radiation of wavelength λ_0 but which is turned on only during a finite interval of time, namely, that time during which the light travels a distance $2l$. Suppose that we receive such radiation in a spectroscope (of infinite resolving power) and want to find out the composition of the spectrum.

Fig. 19

Since we are dealing with an odd function, we shall use equations (58″) and (59″), which yield

$$f(x, 0) = 2i\int_0^\infty A(k) \sin 2\pi kx \, dk,$$

with

$$A(k) = \frac{2}{i}\int_0^l \sin 2\pi k_0 x \sin 2\pi kx \, dx$$
$$= \frac{l}{i}\left[\frac{\sin 2\pi(k_0 - k)l}{2\pi(k_0 - k)l} - \frac{\sin 2\pi(k_0 + k)l}{2\pi(k_0 + k)l}\right].$$

We observe that the function $\left|\frac{\sin u}{u}\right|$ has a maximum ($= 1$) for $u = 0$ and an infinite number of other maxima to the left and right of the former, these other maxima decreasing rapidly as they lie farther and farther away from it. Hence $A(k)$ will be appreciable only for the values of k near k_0, and for these the second term will be negligible compared with the first.[11] The intensity distribution over the spectrum will thus be given approximately by

$$|A(k)|^2 = l^2 \left[\frac{\sin 2\pi(k_0 - k)l}{2\pi(k_0 - k)l} \right]^2, \tag{60}$$

and is graphically represented by the curve of Fig. 20. Neglecting the secondary maxima, we can say that the spectrum is composed of a line which corresponds to the wave number k_0 but which is *broadened:* the width of the line [provisionally defined as the distance between the two minima A and B corresponding to $2\pi(k_0 - k)l = \pm\pi$] is obviously given by $\delta k = 1/l$. If the source were to emit light indefinitely, that is, if instead of a limited *group* of waves we had an unlimited *train* ($l = \infty$), the spectral analysis would yield an infinitely sharp line, as is quite evident.

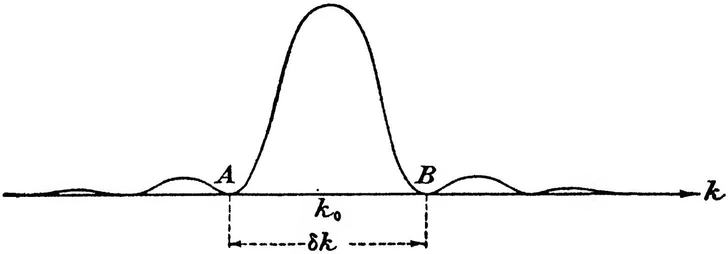

Fig. 20

Therefore we can say that *if a monochromatic wave train is cut short, it loses its monochromaticity, and the spectral line corresponding to it takes on a width inversely proportional to the total length of the wave group* (*or to the duration of emission*).

Hence, the light is no longer strictly monochromatic if it is not emitted over an infinite time interval.

13. Width of a wave group and width of the corresponding spectral line. The observation at the end of the preceding section is of fundamental importance. It may also be extended to wave groups whose amplitude is not constant. (We shall say that we have a group whenever f has appreciable values only within a limited region of the x-axis and is zero or negligible over the remainder.) For this extension, however, we must first give precise and general definitions of the *width* of a spectral line and the *width* of a wave group.

---

[11] Since we are supposed to observe only the "forward"-moving waves, we shall consider only positive values of k.

It is natural, meanwhile, to define as *center* of the spectral line the "center of gravity" of the intensity, that is, the value $\bar{k}$ of k defined b

$$\bar{k} = \frac{1}{I}\int_{-\infty}^{\infty} k|A|^2\,dk, \tag{61}$$

where we have put (total intensity of the spectrum)

$$I = \int_{-\infty}^{\infty} |A|^2\,dk. \tag{62}$$

The half-width Δk of the line will be defined by the following formula (analogous to the one defining the "radius of gyration" in mechanics, or the root-mean-square error in error theory):

$$(\Delta k)^2 = \frac{1}{I}\int_{-\infty}^{\infty} (k - \bar{k})^2|A|^2\,dk. \tag{63}$$

In perfectly analogous fashion, we shall define as *center* of the wave group the point $\bar{x}$ given by

$$\bar{x} = \frac{1}{I}\int_{-\infty}^{\infty} x|f|^2\,dx, \tag{64}$$

where the integral

$$I = \int_{-\infty}^{\infty} |f|^2\,dx \tag{62'}$$

is the same as (62), by virtue of (51′) of §12; and as *half-width*[12] of the group we define Δx given by

$$(\Delta x)^2 = \frac{1}{I}\int_{-\infty}^{\infty} (x - \bar{x})^2|f|^2\,dx. \tag{65}$$

After these preliminaries, it can be shown that the shorter the wave group, the wider will be the spectral line corresponding to it; and more precisely, that Δx and Δk are related by the inequality (of considerable importance in wave mechanics)

$$\Delta k\,\Delta x \geq \frac{1}{4\pi}, \tag{66}$$

which is derived from Fourier's theorem and is independent of the physical significance of the quantities involved.

[12] If this definition were applied, for instance, to the wave group of Fig. 19, it would yield (approximately) $\Delta x = l/\sqrt{3}$; hence the "width" defined as $2\Delta x$ would be $1.155l$.

In order to demonstrate this proposition, let us introduce the function

$$F(x) = f(x, 0)\, e^{-2\pi i \bar{k} x} \tag{67}$$

and let us start from the obvious statement that the quantity

$$D = \left| \frac{x - \bar{x}}{2(\Delta x)^2} F + \frac{\partial F}{\partial x} \right|^2$$

is always ≥ 0. Expanding this expression, we have

$$\begin{aligned} D &= \frac{(x - \bar{x})^2}{4(\Delta x)^4} FF^* + \frac{x - \bar{x}}{2(\Delta x)^2} \left(F \frac{\partial F^*}{\partial x} + F^* \frac{\partial F}{\partial x} \right) + \frac{\partial F}{\partial x} \frac{\partial F^*}{\partial x} \\ &= \frac{(x - \bar{x})^2}{4(\Delta x)^4} FF^* + \frac{1}{2(\Delta x)^2} \frac{\partial}{\partial x} [(x - \bar{x})FF^*] - \frac{FF^*}{2(\Delta x)^2} \\ &\qquad + \frac{\partial}{\partial x} \left(F^* \frac{\partial F}{\partial x} \right) - F^* \frac{\partial^2 F}{\partial x^2} \geq 0. \end{aligned}$$

Let us multiply through by dx and integrate from $-\infty$ to $+\infty$, noting that $FF^* = ff^* = |f|^2$, and taking (65) and (62′) into account. We obtain

$$\begin{aligned} \int_{-\infty}^{\infty} D\, dx = \frac{I}{4(\Delta x)^2} + \frac{1}{2(\Delta x)^2} [(x - \bar{x})FF^*]_{-\infty}^{\infty} - \frac{I}{2(\Delta x)^2} \\ + \left[F^* \frac{\partial F}{\partial x} \right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} F^* \frac{\partial^2 F}{\partial x^2}\, dx \geq 0. \end{aligned}$$

If F is infinitesimal of sufficient order for $x = \pm \infty$, the second and fourth terms will disappear, and there remains

$$- \frac{I}{4(\Delta x)^2} - \int_{-\infty}^{\infty} F^* \frac{\partial^2 F}{\partial x^2}\, dx \geq 0. \tag{68}$$

This integral may be related to Δk as follows. From (63), (59), and (67) we have

$$\begin{aligned} (\Delta k)^2 &= \frac{1}{I} \int_{-\infty}^{\infty} (k - \bar{k})^2 A A^*\, dk \\ &= \frac{1}{I} \int_{-\infty}^{\infty} (k - \bar{k})^2 A\, dk \int_{-\infty}^{\infty} f^*(x, 0)\, e^{2\pi i k x}\, dx \\ &= \frac{1}{I} \int_{-\infty}^{\infty} (k - \bar{k})^2 A\, dk \int_{-\infty}^{\infty} F^*(x)\, e^{2\pi i (k - \bar{k}) x}\, dx; \end{aligned}$$

by reversing the order of integration, we obtain

$$(\Delta k)^2 = \frac{1}{I}\int_{-\infty}^{\infty} F^*(x)\,dx \int_{-\infty}^{\infty} (k - \bar{k})^2 A\, e^{2\pi i(k-\bar{k})x}\,dk. \tag{69}$$

The integral with respect to k may be obtained by noting that (67) may be written, because of (58),

$$F(x) = \int_{-\infty}^{\infty} A\, e^{2\pi i(k-\bar{k})x}\,dk,$$

from which, upon differentiating twice, we get

$$\frac{\partial^2 F}{\partial x^2} = -4\pi^2 \int_{-\infty}^{\infty} (k - \bar{k})^2 A\, e^{2\pi i(k-\bar{k})x}\,dk.$$

Substituting into (69) the expression found in this way by integrating over k, this becomes

$$(\Delta k)^2 = -\frac{1}{4\pi^2 I}\int_{-\infty}^{\infty} F^* \frac{\partial^2 F}{\partial x^2}\,dx.$$

Thus we have obtained the integral that occurs in (68), which therefore becomes

$$-\frac{I}{4(\Delta x)^2} + 4\pi^2 I(\Delta k)^2 \geq 0,$$

from which, since $I > 0$, (66) may be obtained.

The product $\Delta k\,\Delta x$ takes on the minimum value $1/4\pi$ when $D = 0$, that is, when

$$\frac{x - \bar{x}}{2(\Delta x)^2} F + \frac{\partial F}{\partial x} = 0,$$

from which, by integration,

$$F = C\, e^{-\frac{(x-\bar{x})^2}{4(\Delta x)^2}},$$

or else

$$f(x, 0) = C\, e^{-\frac{(x-\bar{x})^2}{4(\Delta x)^2} + 2\pi i\bar{k}x}. \tag{70}$$

This function is represented by a curve of sinusoidal behavior, enveloped by the curve

$$y = C\, e^{-\frac{(x-\bar{x})^2}{4(\Delta x)^2}}, \tag{71}$$

which is the well-known *Gaussian error curve* with its center at $x = \bar{x}$.

This special form of wave group has the peculiarity that $A(k)$ is represented by an expression analogous to f. In fact, upon using (59), we find

$$A(k) = C' \, e^{-\frac{(k-\bar{k})^2}{4(\Delta k)^2} - 2\pi i \bar{x} k}, \tag{72}$$

where

$$C' = \frac{C}{2\sqrt{\pi}\,\Delta k} \, e^{2\pi i \bar{k} \bar{x}}. \tag{72'}$$

Another theorem[13] on the Fourier integral which it is useful to remember concerns the "width" of the wave group and the "width" of the line—defined, however, not in the above fashion but as the smallest interval, which we shall call $2\Delta' x$ (or $2\Delta' k$), containing all the values of x (or of k respectively) for which f (or A respectively) is different from zero. For the case shown in Figs. 19 and 20, we would have $\Delta' x = l$, $\Delta' k = \infty$. The theorem under discussion is the following: *$\Delta' x$ and $\Delta' k$ can never both be finite.* This statement implies, in particular, that if a light source emits light for a finite time, the emitted light will not only be heterochromatic but also its spectrum will (theoretically) be unbounded.

14. Group velocity. Let us now consider an *almost* monochromatic wave group—one such that only frequencies comprised in a small interval appear with appreciable intensities in the Fourier expansion; in other words, $A(k)$ may be considered different from zero only for k lying within a very small interval $(k_0 - \epsilon, k_0 + \epsilon)$. We can then introduce, within such an interval, a variable η, which always takes on very small values, if we put $k = k_0 + \eta$; and hence

$$\nu = \nu_0 + \nu_0' \eta$$

where

$$\nu_0 = \nu(k_0), \qquad \nu_0' = \left(\frac{d\nu}{dk}\right)_{k=k_0}.$$

With this substitution the expansion of f may be written [see (57)]:

$$f(x, t) = e^{2\pi i (k_0 x - \nu_0 t)} \int_{-\epsilon}^{\epsilon} A(k_0 + \eta) \, e^{2\pi i (x - \nu_0' t)\eta} \, d\eta. \tag{73}$$

---

[13] J. von Neumann, *Zeits. f. Physik* **57**, 31 (1929).

Upon comparing this formula with (55) which represents strictly monochromatic waves, we observe that the integral occurring in (73) takes the place of the *amplitude* A of (55). This integral, however, is not a constant but a function of x and t. Therefore it can be stated that equation (73) represents waves of frequency ν_0 and wavelength $1/k_0$, but of variable amplitude (and phase). If for a given t we represent the absolute value of the integral graphically as a function of x (dotted curve in Fig. 21, which we shall call "envelope" of the group), f will be represented by a type of sinusoidal curve of variable amplitude inscribed within that curve.

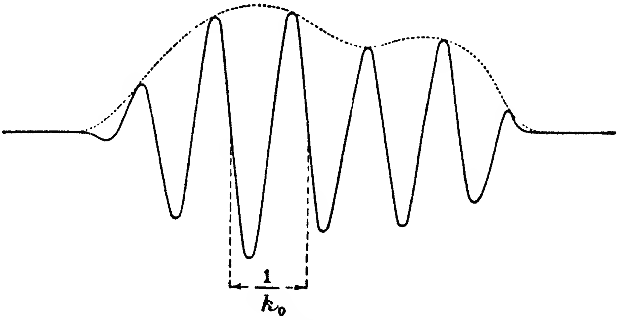

Fig. 21

Then we note that the integral contains x and t only in the combination $x - \nu_0' t$; consequently, the envelope of the group displaces itself without deformation, with velocity ν_0'. Therefore we can say that the entire wave group advances with velocity ν_0', which is called *group velocity* for that reason, and will be designated by v. Thus v will be (implying $k = k_0$)

$$v = \frac{d\nu}{dk}, \tag{74}$$

or else

$$\frac{1}{v} = \frac{d(1/\lambda)}{d\nu}. \tag{74'}$$

It is to be noted that the group velocity v is in general different from the velocity $V = \nu/k$ with which the individual waves travel (*phase velocity*); in fact,

$$\frac{d\nu}{dk} = \frac{d}{dk}[kV(k)] = V(k) + kV'(k)$$

and hence
$$v = V(k_0) + k_0 V'(k_0). \tag{75}$$

Only when $V' = 0$—that is, when there is no dispersion—do we have $v = V(k_0)$.

In order to understand how a wave group may displace itself as a whole, with a velocity different from that of the individual waves, we should observe the waves produced by a pebble on the surface of water. While the group propagates, the outermost waves disappear gradually, while others appear in the interior, so that each of them progresses gradually toward the front of the group, the number of waves within the group remaining constant. The velocity of each wave is therefore larger than the velocity with which the group as a whole travels. This effect is due to the fact that water waves travel with a velocity that is the greater the lower the frequency ($V'(k) < 0$).

15. Waves in space; wave packets.

The statements made in the preceding sections may be easily extended to the propagation of waves in three dimensions. A plane monochromatic wave train can be defined by the wavelength λ, by the direction cosines α, β, γ of the direction of propagation (normal to the wave front), and by the complex amplitude A. The equation which describes this wave is, in analogy to (55), from which it may be obtained by a change of axes,

$$f(x, y, z, t) = A\, e^{\frac{2\pi i}{\lambda}(\alpha x + \beta y + \gamma z - Vt)} \tag{76}$$

or else, upon introduction of the wave number $k = 1/\lambda$ instead of λ, and putting

$$k_x = k\alpha, \qquad k_y = k\beta, \qquad k_z = k\gamma, \tag{77}$$

$$f(x, y, z, t) = A\, e^{2\pi i(k_x x + k_y y + k_z z - \nu t)}. \tag{78}$$

It is convenient to consider k_x, k_y, k_z as components of a vector $\mathbf{k}$ (*propagation vector*) that represents by its absolute value k the wave number $1/\lambda$, and by its direction the direction of propagation.[14] Indicating by $\mathbf{r}$ the vector with components x, y, z which fix the point where f is being considered, one can write (78) in invariant

---

[14] It is useful to remember from now on that in the case of electromagnetic waves this vector is closely related to the momentum $\mathbf{p}$ of the photons (see §3). In fact, $\mathbf{p}$ has the same direction as $\mathbf{k}$ and the magnitude $p = h/\lambda = hk$; therefore

$$\mathbf{k} = \frac{1}{h}\mathbf{p}.$$

form, that is, in a form independent of the axes:

$$f(\mathbf{r}, t) = A\, e^{2\pi i(\mathbf{k}\cdot\mathbf{r}-\nu t)}. \tag{78'}$$

Superimposing such infinite wave trains of all possible wave numbers and directions, we obtain an f represented by

$$f(\mathbf{r},t) = \iiint_{-\infty}^{\infty} A(\mathbf{k})e^{2\pi i(\mathbf{k}\cdot\mathbf{r}-\nu t)}dk_x dk_y dk_z, \tag{79}$$

where ν is in general a function of k_x, k_y, k_z; or of k only, if the medium is isotropic, as we shall assume.

For a given value of t (for example, $t = 0$), this formula may be considered an expansion in Fourier integrals of a given function of x, y, z—an expansion that may be obtained by successively applying the formula of §11 three times. We can see that the coefficient A, the amplitude of the individual wave trains,[15] is given by the following formula, which is the generalization of (59):

$$A(\mathbf{k}) = \iiint_{-\infty}^{\infty} f(r, 0)\, e^{-2\pi i\mathbf{k}\cdot\mathbf{r}}\, dx\, dy\, dz. \tag{80}$$

When f has appreciable values only within a limited region of space and is zero or negligible outside this region, we shall say that it represents a *wave packet* (which is the three-dimensional analogue of the "wave group," defined in §13). A particular wave packet may, for instance, be physically realized in the following manner. Consider a projector capable of emitting a cylindrical pencil of light, which is turned on for a very short time interval, during which the light covers a distance $2l$. There will then originate from the projector a cylindrical "section" of light that constitutes a luminous wave packet.[16]

---

[15] More exactly, $A\, dk_x\, dk_y\, dk_z$ represents the (complex) amplitude of the wave trains having propagation vectors lying between (k_x, k_y, k_z), and $(k_x + dk_x, k_y + dk_y, k_z + dk_z)$.

[16] It is important to note that the wave packet does not rigorously maintain its shape and size unchanged but has a tendency to spread as it propagates. This tendency is caused by diffraction phenomena arising from the diaphragms which limit the light beam, and in addition, if the medium of propagation is not a vacuum, from a possible "dispersion" of the medium. However, within the limits of approximation where geometrical optics holds, we can often ignore this spreading. In this case the packet (which may here be considered to be a point) will evidently describe the path of a "ray," according to the laws of geometrical optics.

As *center* of the packet we define the center of gravity of $|f|^2$, that is, the point whose coordinates $\bar{x}$, $\bar{y}$, $\bar{z}$ are given by

$$\bar{x} = \frac{1}{I}\iiint\limits_{-\infty}^{\infty} x|f|^2\,dx\,dy\,dz \quad \left(\text{with } I = \iiint\limits_{-\infty}^{\infty} |f|^2\,dx\,dy\,dz\right), \tag{64'}$$

and two other analogous formulas. This packet travels with a velocity equal to the *group velocity*, which has already been defined for the one-dimensional case. Similarly one may define, by formulas analogous to (65), the three half-widths Δx, Δy, Δz that give an indication of the extent of the packet along the three axes.

By means of the Fourier expansion, a wave packet may be considered to be obtained by the superposition of infinite monochromatic wave trains, of different propagation vectors, which interfere destructively everywhere except in the region occupied by the packet.

In analogy to (61), we can introduce an *average propagation vector* $\bar{\mathbf{k}}$ defined by

$$\bar{k}_x = \frac{1}{I}\iiint\limits_{-\infty}^{\infty} |A|^2 k_x\,dk_x\,dk_y\,dk_z \tag{81}$$

and by two other analogous formulas; furthermore, we define the quantities Δk_x, Δk_y, Δk_z, in analogy to (63), by

$$(\Delta k_x)^2 = \frac{1}{I}\iiint\limits_{-\infty}^{\infty} (k_x - \bar{k}_x)^2|A|^2\,dk_x\,dk_y\,dk_z, \tag{82}$$

and so on. In (82) we may apply the reasoning of §13 with respect to each of the three variables, finding the inequalities

$$\left.\begin{aligned} \Delta x\,\Delta k_x &\geq \frac{1}{4\pi} \\ \Delta y\,\Delta k_y &\geq \frac{1}{4\pi} \\ \Delta z\,\Delta k_z &\geq \frac{1}{4\pi}. \end{aligned}\right\} \tag{83}$$

These inequalities tell us that *the smaller the extent of the packet in space, the wider must be the distribution of propagation vectors of the* *wave trains composing it.* In practice, the subject of optics deals fairly often with packets of quite large dimensions (compared with the mean wavelength), which may thus be constructed from wave trains of almost a single frequency and direction of propagation.

In a similar way, considering $f(\mathbf{r}, t)$ as a function of t only (that is, fixing our attention upon a definite point of space) and expanding it in Fourier integrals, we would find the relation

$$\Delta\nu \, \Delta t \geq \frac{1}{4\pi}, \tag{84}$$

where $\Delta\nu$ is defined in a manner analogous to Δk_x, and so on, and Δt analogous to Δx, and so on. Therefore, when dealing with a limited wave packet, we can say that Δt measures the half-duration of its travel past a fixed point of space, and $\Delta\nu$ the half-width of the corresponding spectral line, on the frequency scale. In this way we find, from a different point of view, that light of limited duration cannot be monochromatic, and that the shorter the duration, the larger is the width of the spectral line.

16. Note on equations with singularities. In the preceding sections we have always assumed that the coefficients of the differential equation are regular over the entire interval considered, including the end points. In many problems of wave mechanics, however, equations occur in which some of the coefficients of the equation [written in form (1′)] become infinite at some point of the interval, usually at an endpoint (*singular points*). Therefore a brief review will be given here of some of the facts concerning the singularities of linear (homogeneous, second-order) differential equations whose coefficients we shall suppose to be single-valued functions of x.[17]

At a point x_0, the equation (1′) is said to possess a *Fuchsian* (also called *nonessential*) singularity, if for $x = x_0$ at least one of the coefficients P, Q becomes infinite—not to higher than the first order for P, however, and not to higher than the second order for Q—so that the equation may be written

$$y'' + \frac{\mathcal{P}(x - x_0)}{x - x_0} y' + \frac{\mathcal{Q}(x - x_0)}{(x - x_0)^2} y = 0 \tag{85}$$

where $\mathcal{P}(x - x_0)$ and $\mathcal{Q}(x - x_0)$ represent ordinary power series in $x - x_0$ (with integral, nonnegative exponents). If instead for

---

[17] For greater detail see, for example, No. 34 of the Bibliography.

$x = x_0$, the coefficient P becomes infinite to an order higher than the first, or Q to a higher order than the second, the singularity is said to be of the *non-Fuchsian* or *essential* type.

The reason for this distinction lies in the behavior of the solutions in the neighborhood of the point $x = x_0$. In the case of the Fuchsian singularity any integral of the equation is either regular at $x = x_0$ or at most possesses a singularity that permits us to write the solution

$$y = (x - x_0)^\alpha \Phi(x - x_0), \tag{86}$$

α being real (in general not an integer) and Φ being a regular function (or, in an exceptional case which we shall mention below, a function having a logarithmic singularity). On the other hand, if the coefficients at $x = x_0$ do not satisfy the aforesaid condition, there may exist solutions with singularities of different types, or solutions that become altogether indeterminate at $x = x_0$.

Excluding from our considerations the non-Fuchsian singularities, let us start from equation (85) and try to satisfy it by a solution of the form (86), taking for Φ a regular function, that is, a series of integral, positive powers in $x - x_0$, which we may suppose $\neq 0$ for $x = x_0$.

Substituting expression (86) into equation (85) and setting the coefficients of the various powers of $x - x_0$ equal to zero, we may determine formally the coefficients of the series Φ. The first of the conditions obtained in this manner serves to determine α; it is easily found to be

$$\alpha(\alpha - 1) + \alpha\mathcal{P}(0) + \mathcal{Q}(0) = 0. \tag{87}$$

This is a second-degree equation that in general yields two roots, α_1 and α_2, to which there correspond, *in general*, two integrals of the form (86) which are independent (as can be shown) and which thus can be taken as fundamental integrals. The radius of convergence of this series extends to the nearest singular point. If, however, the two roots of (87) differ by an integer (or, in particular, if they coincide), then only one of them (the one whose real part is greater) will furnish a solution of the type (86); the other one gives rise to difficulties in the formation of successive coefficients of Φ. In that case, instead of a series of integral positive powers, we must take for Φ an expression of the type $\Phi_1 + \Phi_2 \log (x - x_0)$, where Φ_1 and

Φ_2 are regular functions. However, we shall not insist upon this exceptional case but shall limit our remarks to stating that in all cases we may find at least one integral of the form (86) by means of (87). Equation (87) is called the *indicial equation* pertaining to the singularity x_0.

Finally, we shall point out the case in which the singularity lies at infinity. This case may be reduced, as is well known, to the previous one by the transformation $x = 1/\zeta$. In this way we can easily find that the necessary and sufficient condition for the point at infinity not to be a singularity of the non-Fuchsian type is as follows: for $x \to \infty$ the coefficient P must be infinitesimal at least of the first order, and the coefficient Q at least of the second order; that is, (for sufficiently large x)

$$P(x) = \frac{1}{x}\mathcal{P}\left(\frac{1}{x}\right), \qquad Q(x) = \frac{1}{x^2}\mathcal{Q}\left(\frac{1}{x}\right), \tag{88}$$

where $\mathcal{P}$ and $\mathcal{Q}$ are positive integral power series in $1/x$.

17. Note on partial differential equations. The equations of interest in wave mechanics are, in most cases, linear, homogeneous partial differential equations. Many of the considerations for the case of a single variable may be extended to such cases. We shall mention each of these extensions briefly, referring here, for convenience of notation, to the case of two independent variables, x and y (the extension to three or more variables is immediate).

The equation will be said to be *self-adjoint* if it has the following form, analogous to (12):

$$\frac{\partial}{\partial x}\left(P\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(P\frac{\partial u}{\partial y}\right) + Ru = 0,$$

where P and R are two functions of x and y (which we shall suppose to be analytic). In R there often occurs a parameter λ, just as in (14); that is, the equation is

$$\frac{\partial}{\partial x}\left(P\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(P\frac{\partial u}{\partial y}\right) + [\lambda + q(x, y)]u = 0. \tag{89}$$

An important particular case is that in which $P = 1$, so that the first two terms form the Laplacian Δu.

The eigenfunction problem, stated for the case of two variables, is the following: *Given a region S of the plane limited by a contour σ, to determine a $u(x, y)$ which within S satisfies* (89) *and which vanishes on the contour, or has a vanishing normal derivative.*

This problem does not in general possess regular nonzero solutions, except for certain values λ_i of the parameter λ, which are called *eigenvalues.* These eigenvalues form (if the region S is of finite extension) a denumerable infinity. To each of them there correspond one or more independent *eigenfunctions* $u_i(x, y)$; if more than one, the eigenvalue is called *multiple.* Two eigenfunctions u_n, u_m, corresponding to two different eigenvalues λ_n, λ_m, have the orthogonality property

$$\int_S u_n u_m^* \, dS = 0 \tag{90}$$

(where the integral is extended over the entire region S, and $dS = dx\, dy$), which is proved in a manner perfectly analogous to the method followed in the case of a single variable.

Similarly we impose upon the eigenfunctions the *normalization* condition:

$$\int_S u_n u_n^* \, dS = 1. \tag{91}$$

There are no generally applicable methods for the solution of problems involving partial derivatives. One that succeeds most often is the method of *separation of variables.* This consists in looking for a solution $u(x, y)$ which is the product of a function X of x alone, times a function Y of y alone. Therefore we must put into (89)

$$u = X(x)Y(y), \tag{92}$$

and we must try, if possible, to effect such a reduction of this equation that we may equate an expression containing functions of x only to one containing functions of y only. If this attempt succeeds, each of the two sides will separately have to be equal to a constant (since both x and y must be able to vary independently). Thus we obtain two equations involving *ordinary derivatives* of the two functions X and Y. Hence the problem is reduced to that of a single variable.

The separation of variables is effectively possible in the majority of the cases in which we are interested here, and we shall see examples later on. It is to be noted that the possibility of separation of variables depends on the coordinate system chosen. By effecting a change of coordinates (for example, from Cartesian to polar coordinates), a separation that was not possible in the original system may sometimes be obtained in the new system. In general, the system of reference that allows the separation of variables has a definite relation to the geometrical properties and to the symmetry of the problem. For example, in problems with spherical symmetry, polar coordinates permit the separation, and so on.[18]

---

[18] See Eisenhart, *Phys. Rev.*, **45**, 428 (1934), and also Pauling and Wilson, *Introduction to Quantum Mechanics*, Appendix IV (McGraw-Hill, 1935).

CHAPTER 6

Probabilistic Statement of Atomic Problems

In this chapter we shall show how the fundamental problems arising from atomic phenomena must be stated, in order not to violate the logical principle which Heisenberg placed at the base of theoretical physics (see §34). That is, quantities which are not at least conceptually observable, or relations which are not at least conceptually verifiable, must no longer enter into our reasoning in an essential capacity. This statement leads, as we shall see, to a resolution of the apparent contradictions between the wave and particle nature of both matter and radiation. However, since the reasoning is perhaps easier to understand in the case of radiation, we shall start with the latter, and shall then transfer our considerations to the case of matter by analogy.

18. Analysis of the photon concept. If we think of the different means which we generally use to detect radiation (vision, photographic plate, photoelectric cell, and others) we recognize that they detect the radiation only by means of the modifications experienced by the material particles absorbing the radiation. These modifications consist essentially in the acquisition of a certain amount of energy (*quantum*) which manifests itself by chemical actions (on the retina or the photographic plate) or thermal actions, or else by the emission of electrons, and so on. We shall call such phenomena *elementary absorption processes.* Another, less common, way to detect radiation consists in making use of the *pressure* which it exerts, or of the *momentum* which it imparts to an electron, as in the Compton effect. In these cases we make use, not of the energy, but of the "momentum" imparted to matter by radiation. Similarly, we may detect that a material particle has completed an *elementary emission process* either by the fact that its energy has decreased without any other cause or by a change in its momentum (recoil) not attributable to any other reason. To sum up, radiation is detected in all cases through its interaction with material particles. It is not possible to detect it on its trajectory from one to the next particle but only at the "departure" or at the "arrival."

For a better understanding of the scope of this impossibility, it is useful to discuss several apparent violations.

A rather trivial one is the well-known fact that when a light beam traverses a turbid medium, its path becomes laterally visible. But it is easily understood that in this case it is only the coarseness of the means of observation which gives us the illusion of seeing a continuous luminous streak. If we observed it with sufficient accuracy, however, we should see the individual scattering particles in the form of separate brilliant points—actually visible to the naked eye in the case of pollen, or with the microscope in the case of turbid solutions. If, then, the scattering particles are single molecules, they are only practically, not conceptually, invisible. But the succession of these brilliant points does not, in fact, outline the points along the trajectory of one and the same photon. Instead, we are clearly dealing with a situation where *some* of the quanta emitted by the source are scattered laterally, while those which are not scattered pass by unobserved. If the source emitted a single quantum, the quantum would be scattered by a material particle, and the observer would see a *single* bright point; or else it would not be scattered, and the observer, standing to one side, would see nothing.

Let us now examine the validity of the following argument. If at A there is a light source which undergoes an elementary emission process, and if at B there is a particle completing an elementary absorption process, it seems possible to assert that the quantum emitted by A *has traveled along the straight line segment* AB in order to reach B where it was absorbed. But if we try to find some means to check this assertion, we shall soon realize its inconsistency. In fact, to lend credence to this assertion, it would be necessary, by placing an opaque screen with a small hole between A and B, to find out if the particle B receives the quantum *only* if the hole lies on the line AB. However, it is well known that a small opening gives rise to diffraction phenomena by virtue of which, if the hole is on the line AB, it *may* happen that the particle B does not receive the light; conversely, it *may* happen that B does receive the light even though the hole is located outside of the line AB.

We might suppose that it is legitimate to attribute to the photon the trajectory constituted by the path ACB when, after we have placed the hole of the screen at any point C and have repeated the experiment many times, it finally happens that particle B receives the energy emitted from A. Evidently we may then repeat the argument given above for each of the portions AC and CB, and we find that the construction of the trajectory with two segments of a straight line is unjustified.

Then we might think that we might achieve the complete determination of the trajectory by a limiting experiment, either by means of a succession of an infinite number of screens each with an infinitely small hole, or, equivalently, by constructing between A and B an infinitely narrow tube with absorbing walls of any shape, rectilinear or not. Then whenever it happens that particle B receives the quantum, we could maintain that the quantum has followed the trajectory defined by the tube. But such a statement would be devoid of physical significance, since we have no other way of checking it than to insert, at an intermediate point of the tube, a material particle capable of absorbing or scattering the light, which would then not reach B any more. This amounts to saying that the phenomenon under study would be altered in a profound way by the very fact of being observed.

We see from these considerations that it is possible to define, in an operational sense, the emission or absorption of a photon but not the trajectory followed by it. Hence it is not permissible to imagine the phenomenon as the motion of a point or of a particle, and it is not very appropriate to speak (as we have done just now, for the sake of simplicity of expression) of a photon's "departure" from or "arrival" at an atom. It is true that the elementary emission process has some analogies to the ejection of a material projectile (recoil and energy loss) and that the elementary absorption process has some analogies to the collision of a projectile with the absorbing atom. Furthermore (for the simple case where no media lie in between), the second phenomenon occurs with a delay with respect to the first, equal to the time required by an object to travel with velocity c. But the analogy does not extend to the intermediate stage of the phenomenon, since the fundamental property of the kinematics of a point is missing, namely, the continuity of the trajectory. Our mental habit makes us consider this property as being inseparably connected with the others, whereas it is logically independent of them. We are led into the error of constructing a *literally* corpuscular model for optical phenomena—to think of photons as ordinary bodies extremely reduced in size but gifted with ordinary kinematic properties. Thus we commit the same error which we should fall heir to if we were to interpret literally the well-known analogies between electric and hydrodynamic currents, taking electricity as an actual liquid. It is not surprising that a *literally* corpuscular theory leads to wrong results, since a premise contrary to experience has been introduced, namely, that of a complete analogy between photons and ordinary bodies.

It is therefore necessary to attribute to the corpuscular model only the value of an analogy limited to certain aspects of phenomena, such as conservation of energy and momentum. If this limitation is kept in mind, the corpuscular model is still of great help, since it furnishes an expressive language for the description of many phenomena by avoiding long circumlocutions. For instance, we shall say "a photon arrives" on a certain surface element $d\sigma$, instead of, "If this surface element were covered with a perfectly absorbing substance, an atom of the latter would undergo an elementary absorption process." Similarly, the expression "to find a photon in a certain volume element dS" means "to fill dS with a completely absorbing material and to determine that an atom of the latter undergoes an elementary absorption process."

In order to see now under what limitations it is permissible to use corpuscular terminology, we shall once again take up the example (see §15) of a projector which, upon being turned on for a short time, emits a cylinder or "packet" of light (which we shall assume to be almost monochromatic). If we want to describe the phenomenon objectively without having recourse to any model, neither wave nor corpuscular, we must say that there is a region S of space (displacing itself with velocity c) which is characterized by the property that, when an atom finds itself within it, the atom may undergo an elementary absorption process. In that case it absorbs a certain quantity of energy w characteristic of the radiation. The probability that the absorption takes place is proportional (for a given type of absorbing atom) to the energy density W, which we assume to be uniform. If we want to translate this situation into the corpuscular model, we shall say that the region S is populated with photons, each having an energy w, so that in each unit volume there will be on the average W/w of them. It is clear that if we take an element of volume dS such that $W/w\ dS$ is small compared to unity, this expression does not represent the number of photons which are to be found in dS, but rather the probability of finding one photon. All this is perfectly identical with what we could say in the case of an aggregate of rubber balls. There is, however, one fundamental difference here. In the case of the balls a perfect knowledge of all conditions of throw would permit us, at least theoretically, to calculate the position of each ball at any instant, and hence to know with certainty whether a ball is to be found in the element dS at a given instant. Thus the probability concept enters into the question only because of the *practical* impossibility of knowing all the elements needed for a precise calculation of the trajectory. On the other hand, in the case of light, there is no meaning in assigning to each photon a position and a velocity within the aggregate, so that no calculation, even in theory, may predict whether a photon is to be found within the element dS at a given instant. All we can say is that the probability of finding it there, when observing it, is $W/w\ dS$. In this case, therefore, the concept of probability enters into the question, not because of our ignorance of some elements of the problem, but because of the nature of the phenomenon itself. Nevertheless, once this important conceptual difference between the two cases has been pointed out, we may formally ignore it at times, or else use, in connection with photons, the same terminology which would be used for material projectiles, at the expense of considering the elements of their motion as *imperfectly known*—or else known with an error whose law of probability is known. Actually, these elements are *physically undetermined*, at least in part, or else *conceptually unobservable*, as will be explained more fully below.

One always has to remember that to say the elements of the motion of a photon (position, trajectory, momentum, and so on) are not "physically determined" has a meaning profoundly different from saying that these elements, though physically determined, are not "known" but could be known a priori through a conceptually possible observation without disturbing the phenomenon. To understand the difference, visualize the following example: Between a source O which emits photons (at time intervals which may be large) and a screen s upon which they fall, there is an opaque screen with two holes A and B. Interference fringes will be formed on the screen. For each photon it is *undetermined* whether it passed through hole A or B. Any attempt to decide this question, as by blocking one of the two paths, would alter the interference effect. Let us now suppose an installation which automatically closes one of the two holes and then the other in an irregular manner, so that the passage of each photon through A or B will depend on an automatic chance law. To an observer who does not see the mechanism, it is *unknown* whether a given photon has passed through A or through B, but contrary to the previous case, it is conceptually possible to decide the question without altering the phenomenon. The difference lies in the fact that in this case, no interference fringes are formed on the screen s. Thus the phenomenon is physically different.

19. Probabilistic meaning of wave optics. We shall now examine the significance which must be attributed to the wave theory of light, and its relation to experimental facts. It is known that all the experiments on image formation, interference, diffraction, and so on, exactly confirm the predictions of the wave theory. These confirmations consist essentially in the following: given the shape, position, and nature of the sources and the various media (lenses, screens, mirrors) we calculate, by means of the laws of optics (or else, more exactly, by means of the wave equation) what the intensity of illumination I should be at various points of the screen, which in general is the retina or a photographic plate. We perform the experiment and find that the measured intensity corresponds, point for point, to the function I coming from calculation and reproducing the predicted shape of the images, shadows, interference fringes, and other factors. In view of the already mentioned discontinuous nature of absorption phenomena, the "measured intensity of illumination," in the case of monochromatic light, to which we shall limit ourselves for now, divided by the energy w of a photon, represents the number of photons per unit area per second arriving on the screen. The verification therefore consists in determining that on an area element $d\sigma$, in a time dt, there arrive $I/w\, d\sigma\, dt$ photons, where I is given as a function of position, by the integration of the wave equations (precisely, I is the projection of the Poynting vector upon the normal to the surface).

It is to be noted, however, that this statement can be verified only if the photons are plentiful (as usually happens in optical experiments). If they are scarce or if there is just one, wave optics breaks down. If, for instance, the source emits a single quantum of energy into a photographic plate, wave optics predicts that the energy must be spread over the whole plate with a certain intensity I, whereas in reality we shall find all the energy at a single point. Indeed, it is just these cases that have suggested the corpuscular model. What significance, then, may be attributed to the intensity I given in wave optics? Evidently it will measure the "probability density" that the photon will arrive at one point of the plate, rather than at another. More exactly, if n photons are emitted, then each emitted photon has the probability $I/nw \, d\sigma \, dt$ of arriving on the element $d\sigma$ in the time dt.

This probabilistic interpretation of I is valid in every case. If the photons are numerous, we can be practically certain, by virtue of the law of large numbers (or Bernoulli's law) that there will actually arrive on $d\sigma$, in the time dt, a number of photons equal to this probability multiplied by n (that is, equal to $I/w \, d\sigma \, dt$) carrying an energy $I \, d\sigma \, dt$. In this way we again find the ordinary meaning of I. Thus we see that *wave optics is merely a method of calculating the probability density of the photon distribution.* When the photons are very numerous, their effective density of distribution, because of the law of large numbers, turns out to be proportional to this probability density, which explains the reason for the agreement of theory and experiment in all ordinary optical phenomena.

Hence we have found that although it is useless to look for laws of motion of photons as they exist for material particles, we have nevertheless a means of determining their distribution statistically. This method is provided by the laws of wave optics.

20. The principle of superposition in optics. We shall now proceed to the case of radiation which is not monochromatic, for example the case of two radiations of frequencies ν_1 and ν_2, with the possibility of their directions and intensities being different as well. Experience shows that the manifestations of these two radiations are *additive;* that is, their effects are superimposed without mutual perturbation. This effect is to be expected in wave theory from the linearity of the differential equations which govern it. Additivity is preserved even in the probabilistic interpretation:

if the waves of frequency ν_1 produce an intensity of illumination I_1 over a certain area, and those of frequency ν_2 an intensity I_2, there is a probability proportional to I_1/w_1 that the element of area $d\sigma$ (assumed to be perfectly absorbing) will absorb a quantum $w_1 = h\nu_1$ in the time dt, and an independent probability proportional to I_2/w_2 that it will absorb a quantum $w_2 = h\nu_2$. This independence of the probabilities, which may be extended to as many monochromatic components of the radiation as desired, is called *principle of superposition* of optics.

We shall add a few words concerning the case in which the *momentum* of the photons is observed, as in the Compton effect. If in wave optics we deal, at a given point, with the passage of plane monochromatic waves having a certain propagation vector $\mathbf{k}$, then all photons observed will have a momentum $\mathbf{p} = h\mathbf{k}$ (see footnote, page 116.) But if, for instance, two plane monochromatic wave trains are superimposed at the point under consideration, having propagation vectors k_1 and k_2 and intensities I_1 and I_2, then the principle of superposition states that there exists a certain probability, proportional to I_1/w_1, of finding a photon with momentum $\mathbf{p}_1 = h\mathbf{k}_1$, and a probability, proportional to I_2/w_2, of finding a photon with momentum $\mathbf{p}_2 = h\mathbf{k}_2$. We find analogous results for the superposition of as many wave trains as we wish, that is, for any radiation.

It will be noted that the conditions described (superposition of wave trains of different frequencies and directions) may be intuitively imagined in the corpuscular model, when the photons are very numerous; it suffices to think of a mixture of particles of different energies and momenta. However, if we are dealing with a single photon or just a few photons, the situation cannot be represented by any classical model. In fact, as long as no experiment is performed to determine the frequency and the momentum of the photon by altering them, it is conceptually impossible to foresee what result such an observation would yield; we can only assign probabilities to the various possible results. Hence the photon does not possess a determined energy and momentum up to the instant of observation; that is, for energy and momentum an *uncertainty* exists which is perfectly analogous to the uncertainty already considered for the position of single photons. Sometimes it may be convenient to have recourse to the corpuscular model in this case as well, replacing this indeterminacy by the supposition that the energy and momentum of the particle are unknown (or better, that only the laws of probability governing them are known). But it should always be remembered that this terminology is inappropriate, as the example of §18 shows.

21. The uncertainty principle for photons. We shall now investigate the limits of validity of the corpuscular model, or else the limits within which it is meaningful to speak of the concepts of "position" and "momentum" of a photon at a certain instant. For this purpose we shall again use the example of the "light packet" S like the one considered in §§15 and 19. Supposing that the source has emitted a single photon, we can say with certainty that the photon is to be found within the region S, constituting the packet (which may be determined by the laws of wave optics). Thus, though it is meaningless to speak of exact coordinates of the photon, we may assign limits to them which are the more restricted the smaller the packet.

We shall now be concerned with the momentum of the photon. From §15, we recall that the wave packet may be considered to be the result of the superposition of an infinite number of monochromatic wave trains of different propagation vectors $\mathbf{k}$, according to the decomposition formula (79) of Chapter 5. By the principle of superposition, there corresponds to each of these wave trains a value $\mathbf{p}$ for the momentum of the photon; that is, to the waves of propagation vector $\mathbf{k}$ there corresponds the possibility of finding the photon with a momentum $\mathbf{p} = h\mathbf{k}$. Hence there is an uncertainty concerning momentum as well as position. To state the matter precisely, if we remember that the amplitude of the wave trains having a propagation vector lying between (k_x, k_y, k_z) and $(k_x + dk_x, k_y + dk_y, k_z + dk_z)$ is $A\,dk_x\,dk_y\,dk_z$, where $A(\mathbf{k})$ is given by formula (80), we easily recognize that the probability that the momentum of the photon lies between (p_x, p_y, p_z) and $(p_x + dp_x, p_y + dp_y, p_z + dp_z)$ is proportional to $|A|^2\,dp_x\,dp_y\,dp_z$. This probability may be expressed in terms of p_x, p_y, p_z rather than of k_x, k_y, k_z by simply inserting $\mathbf{p}$ as a function of $\mathbf{k}$ into (80), which then becomes

$$A\left(\frac{\mathbf{p}}{h}\right) = \iiint_{-\infty}^{\infty} f(\mathbf{r}, 0)\, e^{\frac{2\pi i}{h}\mathbf{p}\cdot\mathbf{r}}\, dx\, dy\, dz. \qquad (80')$$

In order to make this matter more explicit, it is convenient to introduce a space representative of momenta, in which each vector **p** is represented by a point of coordinates p_x, p_y, p_z. We can then say that, just as the position of the photon is uncertain in x, y, z-space and its probability density at different points is proportional to the energy density $W(x, y, z)$ (and hence to $|f|^2$), so the probability density of the momentum in p_x, p_y, p_z-space is represented by the function $|A|^2$ (which similarly defines a wave packet in momentum space). As a measure of the uncertainty in the coordinates, we may take the half-widths Δx, Δy, Δz of the wave packet, defined as in §13. Likewise, the uncertainty in the three momentum components will be measured by the quantities

$$\Delta p_x = h\Delta k_x, \qquad \Delta p_y = h\Delta k_y, \qquad \Delta p_z = h\Delta k_z, \tag{93}$$

where Δk_x, Δk_y, and Δk_z are defined by (82) and two other analogous expressions.

We now recall the theorem proved in §15, which showed that the more restricted the wave packet in x, y, z-space, the greater the difference between the propagation vectors of the waves of which it is composed, or else the wider the packet in p-space. We reach the important conclusion that *the more exactly the position of a photon is determined, the more uncertain the momentum, and vice versa.* More precisely, between the uncertainties Δx, Δy, Δz of the coordinates and Δp_x, Δp_y, Δp_z of the respective momenta the following inequalities hold, which may be deduced immediately from (83) by multiplying through by h:

$$\Delta x\,\Delta p_x \geq \frac{h}{4\pi}, \qquad \Delta y\,\Delta p_y \geq \frac{h}{4\pi}, \qquad \Delta z\,\Delta p_z \geq \frac{h}{4\pi}. \tag{94}$$

Now the partial uncertainty mentioned in §18 is precisely established, and it must be applied to the corpuscular model in order to adapt it to experimental facts. We note that it is not theoretically impossible to localize a photon, at a given instant, with as great a precision as desired. However, the momentum of the photon, and hence its energy and direction of propagation, will then be completely uncertain.

Similarly, if we multiply (84) by h and recall that the energy w

of a photon is given by $h\nu$, we have

$$\Delta w\, \Delta t \geq \frac{h}{4\pi}, \tag{95}$$

which expresses the fact that *the more exactly the instant of passage of a photon through a given point of space is determined, the more uncertain its energy, and vice versa.*

Equations (94) and (95) express the *uncertainty principle for photons.* As may be seen, they are purely mathematical consequences of the fact, proved by experience, that the distribution probability of photons can be calculated from the wave theory.

In many cases it is not necessary to state precisely the value of the right-hand member of (94) and (95), it being sufficient to keep in mind that they are of the order of magnitude of h, and to assume for Δx . . . and Δp_x . . . any measure of the order of magnitude of the uncertainty. The conceptual scope of the principle remains the same. For this reason the uncertainty principle is often expressed by the formulas

$$\Delta x\, \Delta p_x \geq \sim h, \qquad \Delta y\, \Delta p_y \geq \sim h, \qquad \Delta z\, \Delta p_z \geq \sim h, \tag{94'}$$

$$\Delta w\, \Delta t \geq \sim h, \tag{95'}$$

where the symbol $\sim$ indicates that we are dealing with orders of magnitude.

In order to illustrate the principle better and to show the nature of the uncertainty which we are discussing, let us consider a one-dimensional case, limiting ourselves to an estimate of the orders of magnitude. Suppose that we have a group of light waves of total length $2l$, like the group represented in Fig. 19. In such a case, the determination of the momentum $\mathbf{p}$ reduces to a measurement of its absolute value $p = h/\lambda$, that is, to a measurement of the wavelength. For this measurement we imagine the light to fall upon a grating. The grating will make it possible to measure λ with an uncertainty $\delta\lambda$ whose order of magnitude is given (as is known from the theory of the resolving power of gratings) by

$$\frac{\delta\lambda}{\lambda} = \sim \frac{\lambda}{L},$$

where L is the maximum difference in optical paths used by the grating. Evidently, L cannot exceed the length $2l$ of the wave group, and hence

$$\frac{\delta\lambda}{\lambda} \geq \sim \frac{\lambda}{2l};$$

and since we have (ignoring the sign) $\delta p/p = \delta\lambda/\lambda$, the preceding expression yields

$$\frac{\delta p}{p} \geq \sim \frac{\lambda}{2l}, \qquad \text{or else} \qquad \delta p \geq \sim \frac{h}{2l}.$$

This relation between the orders of magnitude of $2l$ and of δp, which could be made more precise by more detailed considerations, simply shows that p will come out the more uncertain, the shorter the wave group, that is the more exactly the position of the photons is determined at any given instant. Of course the argument given with reference to the grating can be repeated for any other means of spectral analysis, because it is based upon the purely analytical fact that the wave group is equivalent in all its effects (as shown in §12) to the superposition of wave trains of different frequencies.

The uncertainty principle expressed by (94) and (95) may be said to represent the corrections which must be applied to the corpuscular model of photons in order to take into account its lack of agreement with reality. Similarly, it can be shown that when the model of electromagnetic waves is used, a certain *uncertainty relation* must be introduced in the measurement of the fields **E** and **H** (see Nos. 10 and 10a of the Bibliography) upon which depends, for instance, the paradox of the photoelectric effect mentioned in §8 of Part I.

22. The uncertainty principle for material particles. The conceptual foundation of quantum mechanics lies in the fact (shown for the first time by Heisenberg) that the uncertainty principle, illustrated in the previous section, holds not only for photons but also for any material "particle," such as electrons, for instance.[1] This is equivalent to saying that *it is conceptually impossible to determine the position and momentum of a particle at a given instant, with a precision exceeding that allowed by the inequalities* (94); just as *it is conceptually impossible to determine the energy of a particle, by means of an experiment lasting a time Δt, with an uncertainty below Δw, related to Δt by* (95).

Of course, this principle does not mean that it is impossible to

---

[1] From this statement it follows that it is improper to use the term "particle" to designate those entities which have (just like photons) only some of the properties which correspond to the ordinary concept of a material particle or corpuscle. However, for convenience we shall currently use this term to collectively designate electrons, protons, neutrons, and nuclei, pointing out once and for all that the term does not have its ordinary meaning.

know the position of the particle to any degree of desired accuracy (just as for photons), or to know its momentum (and hence its velocity) to any wanted precision; but it asserts that these data may not both be known at the same time. More exactly, the principle prohibits the simultaneous knowledge, with infinite precision, of a coordinate and its corresponding component of momentum (or velocity). This uncertainty arises from the inadequacy of the particle model (just as in the case of photons), and marks the limits of validity of that model. This principle, although valid for any body, in practice has important consequences only in connection with particles of atomic or subatomic dimensions, because with ordinary bodies, in view of the smallness of the constant h which occurs in the right-hand member of equations (94) and (95), the uncertainties Δx, Δp_x, and so on, required by the uncertainty principle are negligible compared with those much more considerable uncertainties caused by accidental errors of measurement or the imperfect definition of the reference points.

The fact that the uncertainty principle must also hold for material particles may be logically deduced from the validity of an analogous principle for photons and from the laws of conservation of energy and momentum. These laws, as has been ascertained experimentally, are verified in all phenomena of interaction between material particles and photons. In fact, if the uncertainty principle did not hold for material particles, we could profit by one of these phenomena of interaction to violate the uncertainty principle for photons. It would be sufficient to determine the coordinates of the particle during the interaction with the photon, and its momentum before and after such an interaction, with an accuracy exceeding that corresponding to (94), in order to know, with equal precision, the coordinates of the photon and the momentum imparted to it. For example, in the one-dimensional case mentioned at the end of the preceding section, it would be enough to measure p, not with a grating but by causing the photon to be absorbed by an atom (whose x has also been determined) and by measuring the momentum acquired by the latter. Similarly, given the duration Δt of the interaction, if it were possible to measure the energy of the particle in a time small compared with Δt and with an uncertainty below $h/4\pi\,\Delta t$, we could, by performing this measurement immediately before and after the interaction, determine the energy given up to or taken from the photon with a precision exceeding that allowed by (95).

But to understand the nature of the uncertainty principle, we must examine the mechanism by which it is realized in nature, which is as follows. When we attempt to determine the coordinate x of a material particle with an accuracy Δx, we are forced to impart to it a momentum p_x whose order of magnitude is $h/\Delta x$ but whose precise value it is conceptually impossible to know. Thus if a previous measurement had told us the value of p_x, such knowledge will be partially lost on account of the phenomena taking place after the measurement of x; similarly, a measurement of the momentum component p_x, with an accuracy Δp_x, implies that we displace the particle along the x-axis by a conceptually undetermined amount, of order of magnitude $h/\Delta p_x$, so that, if x was known before this measurement, this knowledge will become partially useless after that instant. Therefore we can say that in order to measure a coordinate and its corresponding momentum (or velocity component), mutually exclusive processes have to be employed. For this reason the two concepts of position and velocity, although they have precise physical meanings separately, may not be attributed simultaneously to the same particle in a precise manner. Bohr has used the term *complementarity*[2] to designate this particular logical relationship between the two concepts.

We shall now show by means of some examples the physical incompatibility of the two types of measurement, referring the reader for a more exhaustive discussion to Heisenberg's book (see Nos. 10 and 10a of the Bibliography).

23. Measurement of the coordinates of a particle. *First method.* In order to determine the position of a particle, we can ideally follow the same procedure ordinarily used to examine bodies, namely, by illuminating the particle and collecting the scattered radiation in an optical instrument (microscope or photographic camera). All that hinders the practical performance of this experiment is the very low intensity of the scattered light, a practical difficulty that in no way negates the validity of the following argument.

Let us suppose that light (or, more generally, radiation) traveling

---

[2] See No. 21 of the Bibliography. We shall see later on that this relation can also be extended to other pairs of physical quantities.

along the x-axis (Fig. 22) strikes the particle whose position we wish to determine. We shall then put up, along the z-axis, an optical instrument, for simplicity assumed to be a pinhole camera without objective, equipped with a hole of center F and radius R. (The argument may also be made by substituting a lens camera or a microscope.) The arrangement will lend itself to an approximate determination of the two coordinates x and y of the particle at a certain instant. In fact, if the particle remained fixed, as at P, and if the experiment took a sufficiently long time, a diffraction pattern would form on the plate L, which we may represent schematically by a small uniform disk of center P' and radius

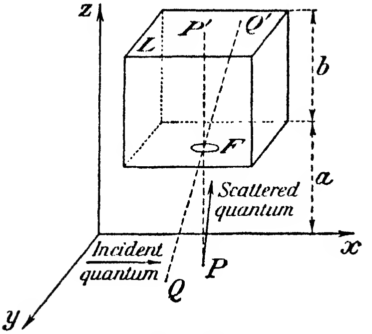

Fig. 22

$$r' = \frac{\lambda b}{R}$$

where b is the distance from the hole to the plate and λ is the wavelength of the scattered radiation. But we know, from the experiments on the Compton effect, that the particle scatters light in quanta and that it receives, in each scattering process, a momentum which alters its velocity. Hence we must limit our experiment to a single scattering process, which will then not yield the whole diffraction pattern on the plate but only a single point Q' of the pattern. The position of the point P', which is where the center of the diffraction pattern would be formed if the particles continued to scatter quanta always from the same position, remains undetermined. All we can say about it is that its distance from Q' will not exceed r'. Hence, if the line $Q'F$ intersects the xy-plane in a point Q, we can say that during the scattering process the particle P was located within a circle of center Q and radius

$$r = \frac{a}{b}r' = \frac{\lambda a}{R}$$

where a is the distance of F from the xy-plane. Thus the coordinates x and y are determined with an uncertainty Δx, Δy of the order of magnitude r:

$$\Delta x = \sim \frac{\lambda a}{R}, \qquad \Delta y = \sim \frac{\lambda a}{R}. \tag{96}$$

(It may be seen that in order to obtain a fairly precise determination of x and y, it is convenient, other conditions being equal, to select radiations of fairly small wavelength, gamma rays for example, rather than light. This observation also corresponds to the well-known fact that the resolving power of optical instruments is inversely proportional to the wavelength of the light used.)

We shall now be concerned with the velocity, or the momentum, with which the particle is left after the experiment. Suppose that initially the particle had a momentum with components p_x^0, p_y^0, p_z^0, which we shall assume exactly known, in order to introduce the most favorable conditions; they may be zero. After the measurement of the coordinates x and y, the momentum p_x, p_y, p_z of the particle will have been changed by the momentum received from the quantum (which we know from the theory of the Compton effect, confirmed by experience). Now the quantum had, before the scattering, a momentum with components

$$\frac{h}{\lambda}, 0, 0. \tag{97}$$

After scattering, if we designate by α, β, and γ the angles formed with the coordinate axes by the direction in which the quantum is scattered, the momentum of the quantum will have components

$$\frac{h}{\lambda}\cos\alpha, \quad \frac{h}{\lambda}\cos\beta, \quad \frac{h}{\lambda}\cos\gamma.$$

Hence the momentum of the particle after scattering will be given by

$$\left.\begin{aligned} p_x &= p_x^0 + \frac{h}{\lambda}(1 - \cos\alpha), \\ p_y &= p_y^0 - \frac{h}{\lambda}\cos\beta, \\ p_z &= p_z^0 - \frac{h}{\lambda}\cos\gamma. \end{aligned}\right\} \tag{98}$$

If α, β, and γ were known, p_x, p_y, and p_z would be determined.

However, as follows from the considerations of §18, it must be observed that the line along which the quantum was scattered cannot be found in any way. We can only say that it will pass through the hole of the pinhole camera, and hence that $\gamma \leq \gamma_0$, calling γ_0 the angle which the radius R of the hole subtends at P, which is given by $\tan \gamma_0 = R/a$; or else, since we are dealing with small angles,

$$\sin \gamma \leq \frac{R}{a}. \tag{99}$$

Then since, evidently,

$$\cos^2 \alpha + \cos^2 \beta = 1 - \cos^2 \gamma = \sin^2 \gamma,$$

we have

$$\cos^2 \alpha + \cos^2 \beta \leq \frac{R^2}{a^2},$$

which is the only limitation on α and β. All we may learn from α and β concerning $\cos \alpha$ and $\cos \beta$ separately, is that they both must be comprised between $-R/a$ and $+R/a$. It follows from this conclusion and from (98) that p_x may vary within the limits

$$p_x^0 + \frac{h}{\lambda}\left(1 \pm \frac{R}{a}\right)$$

and p_y within the limits

$$p_y^0 \pm \frac{h}{\lambda}\frac{R}{a};$$

hence the uncertainty for both is

$$\Delta p_x = \sim \frac{h}{\lambda}\frac{R}{a}, \qquad \Delta p_y = \sim \frac{h}{\lambda}\frac{R}{a}. \tag{100}$$

From these equations and from (96) we obtain, upon multiplication,

$$\Delta x \, \Delta p_x = \sim h, \qquad \Delta y \, \Delta p_y = \sim h,$$

in agreement with (94′). With the camera set up differently, similar reasoning applies concerning the coordinate z and its corresponding momentum.

The same result would be found if, in the case of luminous particles or radioactive particles, we made use of the radiation emitted by the particles rather than of scattered radiation. In fact, the emission of a quantum is accompanied by a recoil which imparts to the particle a momentum in a direction opposite to that of the emitted quantum.

Second method. If a parallel beam of particles (cathode rays, for example) is projected normally against a screen AB (Fig. 23) having a slit of width d, each time a particle crosses the slit we may say that its coordinate y has been determined (assuming the y-axis to lie at right angles to the slit) with an uncertainty

$$\Delta y = \sim d. \qquad (101)$$

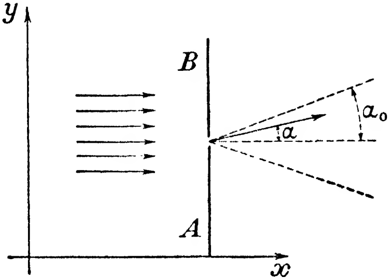

Fig. 23

From experiment (see Chapter 4 of Part I), we may now recall that a beam of material particles of momentum p undergoes diffraction phenomena corresponding to a wavelength $\lambda = h/p$. Therefore beyond the slit the beam will no longer be parallel but will have (in the direction y) an angular width $2\alpha_0$ given, according to the elementary laws of diffraction, by

$$\sin \alpha_0 = \sim \frac{\lambda}{d} = \sim \frac{h}{pd}.$$

That is, the particle, upon crossing the slit, could be deviated through an angle from its original direction, which may go from $-\alpha_0$ to $+\alpha_0$. Therefore the component $p_y = p \sin \alpha$ of its momentum which was originally zero will remain undetermined within the limits $\pm p \sin \alpha_0$, with an uncertainty

$$\Delta p_y = \sim p \sin \alpha_0 = \sim \frac{h}{d}. \qquad (102)$$

From this and equation (101), we now get, in agreement with (94′)

$$\Delta y \, \Delta p_y = \sim h,$$

and we may reason similarly for x and z.

Third method. We know that material particles of large energy, such as alpha particles, may become individually visible, either by means of the scintillations they produce upon hitting a fluorescent screen (spinthariscope), or by being caused to traverse a gas supersaturated with water vapor which condenses, in the form of fog, upon the ions produced by the particle along its path (Wilson cloud chamber). In both these cases it is not actually the particle in question that is localized, but rather the atom which is hit, and is ionized or excited as a consequence of the collision. For this reason the coordinates of the particle at the instant of collision remain undetermined with an uncertainty of the order of the linear dimensions of the atom, which we shall indicate by

$$\Delta x, \quad \Delta y, \quad \Delta z. \tag{103}$$

We shall now see what can be said concerning the momentum which the atom retains after the collision. It is evidently equal to the initial momentum (which we assume to be known) minus the momentum imparted to the electron which is expelled or excited. Thus, in order to determine the latter momentum, we must take the vector difference between the momenta which the atomic electron has before and after the collision. The one after may be measured to any desired accuracy by either of the methods of the following section. However, concerning the momentum which the electron had within the atom before being struck, we can only say that its components lie within certain limits between which they oscillate periodically due to the orbital motion. The amplitude of these oscillations may easily be evaluated in the case of the circular Sommerfeld orbits: it is evidently (choosing the x- and y-axes in the plane of the orbit) given by

$$\Delta p_x = \Delta p_y = mv;$$

and since, by the quantization rule (see §16 of Part I),

$$mvr = n\frac{h}{2\pi},$$

we have

$$\Delta p_x = \Delta p_y = \frac{nh}{2\pi r}.$$

On the other hand, the uncertainty in x and y is given by the dimensions of the orbit in this case, namely,

$$\Delta x = \Delta y = r,$$

so that

$$\Delta x\,\Delta p_x = \Delta y\,\Delta p_y = \frac{nh}{2\pi}.$$

In the most general case of any orbit, we should find a result of the same order of magnitude; that is, in general,

$$\Delta x\, \Delta p_x = \sim nh_y, \qquad \Delta y\, \Delta p_y = \sim nh.$$

Since n in the most favorable case has the value 1, we again find the relations (94′).

24. Measurement of the momentum or the velocity of a particle. *First method.* The most natural procedure of determining the velocity of a particle consists in determining its position at two instants t_1 and t_2, separated by a known time interval. However, we know that every determination of x, y, z introduces an uncertainty in p_x, p_y, p_z, which is given by (94) under the most favorable conditions. Therefore the velocity of the particle after the instant t_2 is not equal to that in the interval between t_1 and t_2, as a result of our measurement. The introduced change remains undetermined, and we know only that its components are of the order of magnitude given by

$$\Delta v_x = \sim \frac{h}{m\Delta x_2}, \qquad \Delta v_y = \sim \frac{h}{m\Delta y_2}, \qquad \Delta v_z = \sim \frac{h}{m\Delta z_2}, \qquad (105)$$

calling Δx_2, Δy_2, Δz_2 the uncertainties which we introduce in the measurement of the coordinates, performed at time t_2. Now, taking $t_2 - t_1$ sufficiently large, we can make the influence of these errors upon the measurement of the ratios $(x_2 - x_1)/(t_2 - t_1)$, and so on, as small as we wish. Hence these ratios may be considered to be exactly determinable, but the quantity of interest, namely, the velocity after t_2, remains affected by an uncertainty expressed by (105), which may be made small only if we are satisfied with little accuracy in the measurement of the position at time t_2. Relations (105), which express this relationship between the two indeterminacies, are just the uncertainty relations (94′).

We observe that the velocity between the two instants t_1 and t_2 (which may be calculated, as we have said, to any desired accuracy) is a quantity devoid of physical interest, because its very definition presupposes that in the interval considered the particle does not interact with its surroundings.

Second method. In order to measure the velocity of a particle without the necessity of recourse to two successive observations of position, we may make use of the Doppler effect. For example, to measure the component v_x we may employ the following scheme.

We send light of a known frequency ν in the direction of the x-axis. We then receive the radiation scattered by the particle in a spectroscope in the negative x-direction and determine its frequency ν'. We then obtain, from an elementary argument of optics,

$$\frac{\nu - \nu'}{\nu} = 2\frac{v_x^0}{c}, \tag{106}$$

where v_x^0 represents the velocity along the x-axis before the measurement. However, it is to be kept in mind that the particle receives a momentum $2h\nu/c$ in the scattering process, and hence that the velocity after the measurement (which is the one we are interested in) is

$$v_x = v_x^0 + 2\frac{h\nu}{mc}.$$

Since ν is supposed to be known, v_x will be determined with the same accuracy as v_x^0 from (106). This accuracy depends on the precision with which ν' is measured; that is, we have

$$\Delta v_x = \Delta v_x^0 = \frac{c}{2\nu}\Delta\nu'.$$

$\Delta\nu'$ cannot lie below $\sim 1/\Delta t$, as we have seen in §15, where Δt is the duration of illumination.[3] Hence under the most favorable conditions we have

$$\Delta v_x = \sim \frac{c}{2\nu\,\Delta t}. \tag{107}$$

On the other hand, we cannot say at what instant of the interval Δt the particle has received the momentum which has altered v_x^0 to v_x. Therefore, after the measurement there remains with the coordinate x of the particle an uncertainty equal to

$$\Delta x = (v_x - v_x^0)\,\Delta t = \frac{2h\nu}{mc}\Delta t.$$

---

[3] Or better, Δt is the duration of a *coherent* wave group. If the illumination lasted longer but suffered sudden phase changes every now and then, Δt would stand for the interval between two such changes.

From this and (107), we obtain

$$\Delta x\, \Delta v_x = \sim \frac{h}{m},$$

or else

$$\Delta x\, \Delta p_x = \sim h,$$

which is the first of relations (94′).

25. Probabilistic statement of the mechanics of a particle. The considerations of the preceding sections show that for material particles as well as for photons, the literally corpuscular representation, which is suggested by the name, does not correspond exactly to reality but merely represents a model whose limit of validity is, so to speak, defined by the uncertainty principle. Therefore we must abandon the intuitive concept of continuous motion along a certain trajectory; instead we must direct our attention toward the spatial distribution of the *probability density*. The latter will be defined, just as for photons, by a function $P(x, y, z, t)$, such that $P\, dx\, dy\, dz$ represents the probability[4] that upon performing an observation at the time t, we find the particle in the element of volume defined by $(x, y, z, x + dx, y + dy, z + dz)$. In analogous fashion, we shall introduce a probability density $Q(v_x, v_y, v_z, t)$ for the components of the velocity of the particle at time t. Hence the mechanics of material particles should not serve to determine their *motion*, that is, their position and velocity as a function of time (which would be contrary to the Heisenberg principle, since it would be equivalent to postulating the existence of phenomena capable of exactly defining the position and velocity of a particle at a given instant). Instead the mechanics of material particles should play a role similar to that which optics plays in relation to photons, namely, to determine the probability densities $P(x, y, z, t)$ and

---

[4] In order to describe precisely the significance which must be attributed to the term "probability" in wave mechanics, we should imagine a *large* number N of independent identical systems, subject to the same initial conditions, upon each of which we perform an observation of the particle at time t. If in N' systems the measurement yields the considered result, we say that its "probability" is N'/N. (In the optical case, we may refer to the *simultaneous* presence of N photons; this is not possible here, because N particles would interact and alter their respective probability distributions.) Similarly, when we speak of the *average value* of a quantity, we understand that this value is to be measured in the N above-mentioned systems and that the average of the results is to be taken.

$Q(v_x, v_y, v_z, t)$. If we then keep in mind that the diffraction phenomena, which constitute some of the most obvious proofs of the wave character of the laws of optics, are also verified for material particles (as we have seen in Chapter 4 of Part I), we are naturally led to inquire whether in the case of material particles the probability densities might not also be obtained from equations of the type used for waves.[5] This idea may be more exactly stated as follows.[6]

In the case of optics, the energy density W (and similarly the intensity I of illumination of a surface) does not obey a simple equation but is obtained from the formula

$$W = \frac{\epsilon}{8\pi}(E_x^2 + E_y^2 + E_z^2) + \frac{\mu}{8\pi}(H_x^2 + H_y^2 + H_z^2),$$

in terms of the components of the electric field **E** and the magnetic field **H**, each of which satisfies the wave equation, which for E_x is

$$\Delta E_x = \frac{n^2}{c^2}\frac{\partial^2 E_x}{\partial t^2} \tag{108}$$

where $n = \sqrt{\epsilon\mu}$, the index of refraction. Similarly[7] in the case of

---

[5] On the limits of the analogy between optics and mechanics, see W. Pauli, *Zeits. f. Physik* **80**, 573 *ff* (1933).

[6] It goes without saying that the heuristic approach used here does not reproduce the historical development of the theory (for which we refer to what was said in Part I), nor does it pretend to give a rigorous justification for the latter.

[7] In order to make the analogy more evident formally, it is convenient to represent the electromagnetic field by a single complex vector $\mathbf{\Psi}$, defined by

$$\mathbf{\Psi} = \sqrt{\frac{\epsilon}{8\pi}}\mathbf{E} + i\sqrt{\frac{\mu}{8\pi}}\mathbf{H}.$$

We can then see immediately that the expression for W becomes

$$W = |\Psi_x|^2 + |\Psi_y|^2 + |\Psi_z|^2 = |\mathbf{\Psi}|^2.$$

The vector $\mathbf{\Psi}$ (at points free of charges and currents) satisfies the equations

$$\text{curl}\ \mathbf{\Psi} = \frac{i}{c}n\frac{\partial \mathbf{\Psi}}{\partial t}, \qquad \text{div}\ \mathbf{\Psi} = 0,$$

which condense the equations of Maxwell and Laplace for **E** and **H**. From these we find immediately that each complex component of $\mathbf{\Psi}$ satisfies the mechanics, rather than to look for a differential equation which P satisfies, it is convenient to introduce a certain number N (undetermined for the present) of scalar, generally complex quantities ψ_1, ψ_2, . . . ψ_N, from which P may be obtained by the formula

$$P = |\psi_1|^2 + |\psi_2|^2 + \cdots + |\psi_N|^2,$$

each of which satisfies a differential equation of the second order, of the wave type. Dirac has shown that $N = 4$ is the minimum number of functions ψ with which a wave mechanics in agreement with experimental facts and satisfying the principle of relativity can be constructed. If, however, we are satisfied here with a wave mechanics that is valid with the approximation with which ordinary nonrelativistic mechanics holds (or for velocities small compared to c), two functions ψ will suffice (Pauli theory—see Chapter 14). If the *spin* effects are also negligible (see §25 of Part I), a single function ψ will be sufficient; this is the wave mechanics of Schrödinger, with which we shall be concerned until Chapter 14, where the relativistic wave mechanics of Dirac with four components ψ will be developed. We shall further limit ourselves here to the study of a single particle.

We shall thus introduce a complex function $\psi(x, y, z, t)$ which Schrödinger called "field scalar" and which is now more generally called *probability amplitude,* in analogy with the amplitude of light waves. The latter has no immediate physical significance but serves to determine the probability density P (susceptible of experimental check, as we have seen) by means of the relation

$$P = |\psi|^2. \tag{109}$$

We shall then assume that ψ satisfies a differential equation analogous to (106), that is,[8]

$$\Delta\psi = N^2 \frac{\partial^2\psi}{\partial t^2} \tag{108'}$$

---

wave equation

$$\Delta\Psi_x = \frac{n^2}{c^2}\frac{\partial^2\Psi_x}{\partial t^2}, \text{ and so on.}$$

[8] This equation holds rigorously only for "monochromatic" waves, that is, waves of a single frequency. However, the principle of superposition enables us to pass quickly to the more general case, as will be seen more clearly in §29.

where N, which represents the reciprocal of the phase velocity and is analogous to n/c of (108), will be determined, by a method explained later, as a function of x, y, z. ψ will thus propagate as waves (de Broglie waves), which, however, have no physical existence but merely provide an analytical means of calculating the probability density P.

Finally, we point out that the integral of P extended over all space expresses the total probability that the particle is to be found anywhere and hence must result as unity. Therefore ψ must satisfy the condition

$$\iiint |\psi|^2 \, dx \, dy \, dz = 1. \tag{110}$$

Since ψ is generally complex, it may be written in the form $\psi = \psi' + i\psi''$, where ψ' and ψ'' are two real functions satisfying (108) separately, since the coefficients of the latter are real. The use of imaginary quantities serves only to simplify the writing, permitting the two equations which ψ' and ψ'' satisfy to be collected into a single formula. It might be supposed that a wave mechanics with a single real ψ could be constructed; but this has been proved impossible[9] if we want P always to come out > 0, as it must to satisfy (110), and if it is to be obtained in finite terms from ψ without the intervention of the derivatives of ψ with respect to t.

It remains to be seen how this probabilistic interpretation of mechanics is connected with the ordinary concepts of kinematics and dynamics, which, as is well known, may in many cases be applied successfully (at least within the limits of observational errors) to electrons, atoms, and so on. These are the cases in which the probability density P is different from zero only in a region so restricted that it may be considered to be a point (or, when the de Broglie waves constitute a small "packet," analogous to the light packet already considered several times). The particle may then be thought of practically as being localized in the center of the packet (or at any point of the packet), and its motion will be identified with the motion of that point. Hence every time we reason in terms of the classical kinematics of the motion of a particle, it is to be understood that we speak of the motion of a packet of de Broglie waves which is practically a point.

---

[9] Pauli, *Zeits. f. Physik* **80**, 573 *ff* (1933).

26. Motion of a particle and motion of a wave packet. In order to give a definite form to the fundamental equation of wave mechanics (108′), the coefficient N, which we may call the *index of refraction* of space for the de Broglie waves, remains to be determined. In general it will be a function not only of x, y, z but also of the frequency ν of these waves (as in optics for dispersive media). To determine N, we must also take into account two experimental facts:

(I) the fact that, in cases in which the de Broglie waves constitute almost a point packet, the motion of the latter is governed by the ordinary laws of point mechanics (see the end of the preceding section);

(II) the relation $\lambda = h/p$ (given by diffraction experiments) between the de Broglie wavelength and the momentum of the particles (see §33 of Part I).

In order to account for condition (I), we shall consider a particle of mass m moving in a field of potential[10] $U(x, y, z)$, with total energy E. In wave mechanics this particle (supposing that its position is known with very great precision) will be represented by a "de Broglie wave packet" which is so restricted in extension that it may be considered to be a point. The movement of this packet between any two points A and B of its trajectory must be identified (ignoring diffraction) either by its trajectory or by its velocity, with the motion which classical mechanics assigns to a material point of mass m and energy E, in a potential U between points A and B. In order to obtain this identification, we have at our disposal the coefficient N of (108′) as well as the frequency ν of the de Broglie waves (or better, the mean frequency of the wave trains forming the packet).

First we shall proceed to the identification of the two trajectories. The trajectory of the wave packet is nothing but a "ray" and is therefore determined by the laws of refraction, which may be condensed into the variational principle. In the case of light waves this principle takes the name of *Fermat's principle* and may be stated as follows: The line s traversed by a light ray between any

---

[10] We designate as "potential" the potential energy of the particle, whereas in rational mechanics "potential" is usually the same energy with its sign reversed.

two points A and B is such that, when it is deformed infinitesimally (keeping A and B fixed) the time needed by the light to travel from A to B does not change, at least to infinitesimals of higher order.

Since the time required by light to cover an infinitesimal distance ds is equal to $N/c\,ds$ (N is the index of refraction, and hence c/N is the velocity of light at the point in question), the principle just stated may be expressed (using the well-known symbol δ from the calculus of variations) by the formula

$$\delta \int_A^B N\,ds = 0. \tag{111}$$

On the other hand, the trajectory of a material point, according to classical mechanics, is determined from the *principle of least action*, which in the present case may be expressed as follows: The trajectory s is such that when it is deformed by an infinitesimal amount (keeping A and B and the energy E fixed) the integral

$$\int_A^B \sqrt{E - U}\,ds \tag{112}$$

(twice this integral is called "action") does not vary, at least to infinitesimals of higher order. This is equivalent to saying that

$$\delta \int_A^B \sqrt{E - U}\,ds = 0. \tag{113}$$

If now the trajectory of the wave packet between A and B is to coincide with the trajectory which classical mechanics ascribes to the point, it is necessary that (111) and (113) be equivalent, for which it suffices to take

$$N = C\sqrt{E - U}, \tag{114}$$

where C is an arbitrary constant (independent of x, y, z). Thus we have found the spatial distribution of the index of refraction (it was to be expected that the latter is a function of x, y, z through U only). We observe that, since in general the index of refraction of de Broglie waves depends on their frequency, C and E must be considered to be functions of ν.

Let us now also identify the velocity of the two motions point for point. The velocity v of the wave packet is not the phase velocity $1/N$ but the group velocity (see §14), which is given by

(74′) and may be written (noting that $1/\lambda = N\nu$) as

$$\frac{1}{v} = \frac{d(N\nu)}{d\nu}, \tag{115}$$

or else, introducing (114),

$$\frac{1}{v} = \frac{d}{d\nu}(C\nu\sqrt{E - U}) = \frac{d(C\nu)}{d\nu}\sqrt{E - U} + \frac{C\nu}{2\sqrt{E - U}}\frac{dE}{d\nu}. \tag{116}$$

On the other hand, the velocity of the point given by classical mechanics is (since $\frac{1}{2}mv^2 = E - U$)

$$v = \sqrt{\frac{2}{m}}\sqrt{E - U}. \tag{117}$$

Equating the two expressions for $1/v$, we obtain, after obvious reductions,

$$\sqrt{2m} = 2\frac{d(C\nu)}{d\nu}(E - U) + C\nu\frac{dE}{d\nu}. \tag{118}$$

We observe that (118) must be an identity in x, y, z, and that these variables enter only through U. The coefficient of U must therefore be identically zero; that is, we must have

$$\frac{d(C\nu)}{d\nu} = 0,$$

or

$$C\nu = K, \tag{119}$$

K being an arbitrary constant. Then (118) becomes

$$\sqrt{2m} = K\frac{dE}{d\nu}$$

or

$$\frac{dE}{d\nu} = a,$$

where we have put

$$a = \frac{\sqrt{2m}}{K}. \tag{120}$$

Upon integrating, we get for $E(\nu)$ the linear expression

$$E = a\nu + b, \tag{121}$$

where a and b are two constants. The expression for $C(\nu)$ is then given by (119), which may be written, by making use of (120),

$$C = \frac{\sqrt{2m}}{a\nu}. \tag{122}$$

Precisely, the expression for the index of refraction which ensures the complete identity of the two motions is

$$N = \frac{\sqrt{2m}}{a\nu}\sqrt{E - U}, \tag{123}$$

where the expression (121) may be substituted for E.

The phase velocity of the de Broglie waves is therefore

$$V = \frac{1}{N} = \frac{a\nu}{\sqrt{2m}}\frac{1}{\sqrt{E - U}}, \tag{123'}$$

and their wavelength is

$$\lambda = \frac{V}{\nu} = \frac{1}{N\nu} = \frac{a}{\sqrt{2m(E - U)}},$$

or else, upon introducing the momentum $p = mv$ and considering (117),

$$\lambda = \frac{a}{mv} = \frac{a}{p}. \tag{124}$$

27. The Schrödinger equation. The expression for N found in the preceding section, when substituted into (108′), furnishes a wave mechanics which satisfies the condition that it coincide with ordinary mechanics within limits analogous to those of geometrical optics, for any constants a and b. But condition (II) of the previous section forces us to fix the value of the constant a. In fact, equation (124) has very nearly the form of the relationship found experimentally between λ and p, but for it to agree numerically we must take

$$a = h. \tag{125}$$

The constant b, however, remains undetermined, since it does not occur in the expression for λ but only in the expression for ν, and the frequency of the de Broglie waves is not accessible to experiment. This condition accounts for the fact that the energy E is defined to within an arbitrary additive constant.[11] For sim-

---

[11] This statement holds only in nonrelativistic mechanics. Taking relativity into account, we have then a way of fixing E in absolute magnitude also; and analogously, in relativistic wave mechanics b is determined (see Chapter 14).

plicity we shall assume that $b = 0$, so that (121) becomes

$$E = h\nu, \tag{121'}$$

that is, identical to the formula which holds for photons.

If we take (125) into account, the relation (123′) between the phase velocity V of the de Broglie waves of frequency ν and the potential U becomes

$$\frac{1}{N} = V = \frac{h\nu}{\sqrt{2m}} \frac{1}{\sqrt{E - U}}, \tag{126}$$

or also from (121′)

$$\frac{1}{N} = V = \frac{h\nu}{\sqrt{2m}} \frac{1}{\sqrt{h\nu - U}}. \tag{126'}$$

Therefore we can say that for de Broglie waves a force field represents the analogue of a medium of nonuniform index of refraction for light, and furthermore that for the de Broglie waves the medium is always dispersive, since the index of refraction depends on ν.

If we now introduce the expression (126) for N into the general equation (108′) which ψ satisfies, we find

$$\Delta\psi = \frac{2m}{h^2\nu^2} (E - U) \frac{\partial^2\psi}{\partial t^2}, \tag{127}$$

which expresses the law of propagation for ψ and is the fundamental equation of wave mechanics. However, (127) presupposes that the waves are *monochromatic* (to use optical terminology), or that ψ is a sinusoidal function of time with frequency ν, which means physically, as (121′) shows, that the energy of the particle has a definite value E (*stationary states* or *quantum states*). In order to express this fact analytically, it is convenient (according to the established use in other branches of physics, such as in electricity) to make use of exponential functions with imaginary exponent, and to put[12]

$$\psi = u(x, y, z)\, e^{-2\pi i\nu t}, \tag{128}$$

where u is a generally complex function independent of time, whose modulus represents the amplitude of the oscillations of ψ, and where $\nu = E/h$.

---

[12] Many authors write this formula with a plus sign in the exponent. This makes no essential difference, except for a few changes of sign in the formulas derived from it, as in the right-hand member of (136).

From (128) we obtain

$$\frac{\partial \psi}{\partial t} = -2\pi i \nu \psi, \tag{129}$$

$$\frac{\partial^2 \psi}{\partial t^2} = -4\pi^2 \nu^2 \psi. \tag{130}$$

By taking these into account, (127) may be written in the following form, which no longer contains derivatives with respect to time, and is the one usually adopted (*Schrödinger equation*):

$$\Delta \psi + \frac{8\pi^2 m}{h^2} (E - U)\psi = 0; \tag{131}$$

or also, upon putting for ψ expression (128),

$$\Delta u + \frac{8\pi^2 m}{h^2} (E - U)u = 0. \tag{131'}$$

This shows that the spatial part u of the function ψ satisfies the same equation as ψ. Since, on the other hand, $|\psi|^2$ determines the probability distribution, which is equal to $|u|^2$, in many cases it makes no difference whether we use the function u instead of ψ (provided, however, that we are dealing, as we are assuming here, with waves of a single frequency).

The function conjugate to ψ, or

$$\psi^* = u^*(x, y, z)\, e^{+2\pi i \nu t}$$

evidently satisfies the same equation (131) as ψ and has the same modulus, so that nothing new is learned by considering it.

It is also to be noted that we are dealing with homogeneous equations in (127) or (131); hence, after one solution has been found, an infinite number of others may be obtained from it by multiplying by an arbitrary constant. However, this indeterminacy is eliminated if we take into account the fact that ψ must satisfy condition (110), as has already been stated in §25. This equation may also be written (setting $dS = dx\, dy\, dz$ as usual) in the form

$$\int \psi \psi^* \, dS = 1. \tag{132}$$

This is called the *normalization condition* and has been studied in Chapter 1 for the eigenfunctions of differential equations. As we have seen in §4, it may always be satisfied (provided that the integral is convergent, a condition which we shall study later on) and determines (to within a factor of the form $e^{i\theta}$) the arbitrary multiplicative constant. The arbitrariness which remains in the *argument* of ψ because of the arbitrary constant θ has no influence either on the modulus or on the wavelength of ψ, which alone have physical significance; hence it may be neglected.

28. Energy levels as eigenvalues of the Schrödinger equation. In order to make the problem of the wave mechanics of a particle determinate, we must impose some conditions of regularity on ψ, in addition to the condition of satisfying (in the case of stationary states) equation (131) with an appropriate value for the constant E (which occurs in the equation as an a priori undetermined constant). Specifically, it will first of all be required that ψ and its first derivatives be continuous and single-valued over all space. Furthermore, in order for us to be able to apply the normalization condition (132), the integral of $|\psi|^2$ extended over all space[13] must be convergent; (this condition will be assured if ψ is infinitesimal to sufficiently high order at infinity. If then the potential $U(x, y, z)$ has points of singularity, singularities for ψ may also occur (see §16). In this case we shall require that these singularities be at most poles of order lower than the first.[14]

With these conditions, the problem of the integration of the Schrödinger equation enters into the category of those studied in the mathematical introduction. As has been pointed out, this equation has solutions only if the parameter E (which corresponds to λ of Chapter 5) has one of the values which we have termed *eigenvalues* of the differential equation. Thus the rather natural conditions of regularity imposed upon ψ lead automatically to the existence of *energy levels* (discrete or sometimes continuous) which is an experimental fact, as we have seen. The determination of these levels reduces to the mathematical problem of looking for the eigenvalues of the Schrödinger equation. Herein lies one of the most remarkable results of wave mechanics, as has been pointed out in Part I.

[13] In certain cases the conditions of the problem force the particle to stay within a certain region S. Then, of course, we may integrate the equation only within that region, with the condition that $\psi = 0$ outside; (see §38, for example).

[14] For a justification of these conditions as well as for a statement of the conditions which ψ must satisfy in more general problems than these, see von Neumann, *Götting. Nachr.*, 1 (1927); also Pauli, No. 14 of the Bibliography, page 121.

28a. The eigenfunctions of the Schrödinger equation. To every eigenvalue E_n (where n may in general represent a group of several indices) there will correspond one or more normalized eigenfunctions which, naturally, depend upon the time according to the law

$$\psi_n = u_n e^{-\frac{2\pi i}{h}E_n t}. \tag{128'}$$

If the eigenvalue is simple, the eigenfunction corresponding to it immediately gives, by the square of its modulus, the probability distribution P. If the eigenvalue is multiple, this distribution will not be completely determined by the value of the energy: this is the case of *degeneracy*, of which we shall see important examples in what is to follow.

We shall now prove an important property of the eigenfunctions of the Schrödinger equation, namely: if ψ_n and ψ_k are two eigenfunctions belonging to two *distinct* eigenvalues E_n and E_k, they are "orthogonal" to each other, which amounts to saying that

$$\int \psi_n \psi_k^* \, dS = 0, \tag{α}$$

where the integral is understood to be extended over all space. This theorem and terminology are the natural extension of what has been said in §5 for the one-dimensional case.

In order to prove (α), we shall first of all write that ψ_n and ψ_k satisfy the Schrödinger equation for $E = E_n$ and $E = E_k$, respectively:

$$\left.\begin{aligned} \Delta\psi_n + \frac{8\pi^2 m}{h^2}(E_n - U)\psi_n &= 0, \\ \Delta\psi_k + \frac{8\pi^2 m}{h^2}(E_k - U)\psi_k &= 0. \end{aligned}\right\} \tag{β}$$

The second equation, when i is changed into $-i$, may be written

$$\Delta\psi_k^* + \frac{8\pi^2 m}{h^2}(E_k - U)\psi_k^* = 0. \tag{γ}$$

Let us now multiply (β) by $\psi_k^* \, dS$ and (γ) by $\psi_n \, dS$, subtract one from the other, and integrate over any volume S'. We obtain

$$\int_{S'} (\psi_k^* \Delta\psi_n - \psi_n \Delta\psi_k^*) \, dS + \frac{8\pi^2 m}{h^2}(E_n - E_k) \int_{S'} \psi_n \psi_k^* \, dS = 0.$$

Let us transform the first integral by means of Green's theorem[15] into an integral extended over the surface σ bounding S' (whose direction-normal pointing into S' we shall call ν). We then get

$$\int \left(\psi_k^* \frac{d\psi_n}{d\nu} - \psi_n \frac{d\psi_k^*}{d\nu}\right) d\sigma + \frac{8\pi^2 m}{h^2}(E_n - E_k)\int_{S'} \psi_n\psi_k^*\, dS' = 0.$$

Let us now make the surface σ tend toward infinity. By virtue of the hypotheses made in §28 concerning the behavior of the eigenfunctions at infinity, the integral extended over σ approaches zero, whereas the integral extended over S' has for limit the integral occurring in (α). Hence there remains

$$(E_n - E_k)\int \psi_n\psi_k^*\, dS = 0,$$

and since $E_n \neq E_k$ has been assumed, relation (α) is proved.

29. The principle of superposition of wave mechanics. It is natural to assume that a principle of superposition analogous to the one in optics holds for de Broglie waves. Let us therefore consider two different (normalized and independent) solutions ψ_1, ψ_2 of the Schrödinger equation, corresponding in general[16] to two different values E_1 and E_2 of the energy and to two different momenta $\mathbf{p}_1$ and $\mathbf{p}_2$. For simplicity we shall limit ourselves to the case of only two components, but the extension to any number is immediate. We now form a linear combination (with two constant coefficients c_1 and c_2)

$$\psi = c_1\psi_1 + c_2\psi_2 = c_1 u_1\, e^{-\frac{2\pi i}{h}E_1 t} + c_2 u_2\, e^{-\frac{2\pi i}{h}E_2 t},$$

to which we shall apply the normalization condition. This becomes $|c_1|^2 + |c_2|^2 = 1$, as may be verified immediately if it is remembered that ψ_1 and ψ_2 are orthonormal. What will be the significance of this ψ? It is evident that this case is analogous to the optical case (§20) of a photon with undetermined energy and momentum, and that therefore $|c_1\psi_1|^2\, dS$ represents the probability that the particle will be found in the volume element dS with energy E_1 and momentum $\mathbf{p}_1$, that is, with the characteristics corresponding to the ψ_1

---

[15] See §9 of No. 32a of the Bibliography.

[16] In cases of degeneracy the two energies E_1 and E_2 may be equal. In that case too, however, we shall require ψ_1 and ψ_2 to be orthogonal (see §6). This case corresponds, in optics, to the superposition of waves of the same frequency but of different directions.

waves, whereas $|c_2\psi_2|^2\,dS$ represents the probability of finding it in dS with the characteristics of the ψ_2 waves. It is to be noted that these two expressions represent the actual probability and not quantities proportional to it, as may be immediately verified by observing that the total probability of finding the particle anywhere and with any energy turns out to be unity (as it should). In fact,

$$\int|c_1\psi_1|^2\,dS + \int|c_2\psi_2|^2\,dS = |c_1|^2 + |c_2|^2 = 1.$$

It is now evident that $|c_1|^2$ represents the probability that the particle has an energy E_1 and a momentum $\mathbf{p}_1$, and $|c_2|^2$ the probability that it has energy E_2 and momentum $\mathbf{p}_2$.

Let us now consider the most general $\psi(x, y, z, t)$ possible. Once t has been fixed at a certain value, $t = 0$ for example, ψ may be expanded in a series by means of the eigenfunctions of the Schrödinger equation (131′), which, as we know, form a complete orthogonal set; hence

$$\psi(x, y, z, 0) = \sum_n c_n u_n(x, y, z),$$

with

$$c_n = \int\psi(x, y, z, 0)u_n^*\,dS.$$

Now each of the u_n represents the distribution of ψ for $t = 0$, in a "stationary state"; this ψ then evolves in time according to (128′). These various components, as we have said, always superimpose themselves without affecting each other, and hence ψ at any later time t will be given by

$$\psi(x, y, z, t) = \sum_n c_n u_n(x, y, z)\, e^{-\frac{2\pi i}{h}E_n t}. \tag{133}$$

This ψ represents a *state*[17] of the particle in which the energy and momentum are not determined: the probability that the energy has a value E_n and the momentum a value corresponding to the waves ψ_n is $|c_n|^2$, because of what has been said above.

Equation (133) has been written for the case of discrete eigenvalues. However, if the eigenvalues of (131′) constitute a continuous spectrum from $E = E_a$ to $E = E_b$, the series will be replaced

---

[17] The form of the function ψ which is assigned to each case (or "state" of the system) depends on the initial conditions (as we shall see better in Chapter 11) and in particular on the observations to which the system has been subjected initially.

by an integral:

$$\psi = \int_{E_A}^{E_B} c(E)\psi_E \, dE, \tag{133'}$$

where $\psi_E = u_E(x, y, z) \, e^{-\frac{2\pi i}{h}Et}$, and u_E satisfies the Schrödinger equation

$$\Delta u_E + \frac{8\pi^2 m}{h^2}(E - U)u_E = 0.$$

In this case, $|c(E)|^2$ represents the "probability density" in the continuous energy spectrum; that is, $|c(E)|^2 \, dE$ is the probability that the energy lies between E and $E + dE$.

The most general case is that in which there are both discrete and continuous eigenvalues, where ψ will be the sum of a series and an integral. Nevertheless, we may consider (133′) to be the most general form of ψ, if the integral is understood to be defined in the sense of Stieltjes (see footnote on page 104).

30. The time-dependent Schrödinger equation. It is to be kept in mind that ψ in the general form (133) or (133′)—that is, the "heterochromatic" form—does not satisfy the Schrödinger equation (131), because each of its terms satisfies an equation of the form (131) with a different value of E. However, we can easily find an equation independent of E which is satisfied by all monochromatic components, and therefore also by the ψ which results from their superposition.

We recall that the component ψ_n is of the form

$$\psi_n = u_n(x, y, z) \, e^{-\frac{2\pi i}{h}E_n t}$$

and therefore

$$\frac{\partial \psi_n}{\partial t} = -\frac{2\pi i}{h} E_n \psi_n. \tag{134}$$

The latter satisfies the Schrödinger equation

$$\Delta \psi_n + \frac{8\pi^2 m}{h^2}(E_n - U)\psi_n = 0. \tag{135}$$

Upon eliminating E_n between (134) and (135), we obtain the equation

$$\Delta \psi_n - \frac{8\pi^2 m}{h^2} U \psi_n = -\frac{4\pi i m}{h} \frac{\partial \psi_n}{\partial t},$$

in which the coefficients are independent of the index n. Therefore this equation is satisfied by all components ψ_n, and hence also by any of their linear combinations.

We may now say that the probability amplitude ψ, even in the case where the energy of the particle is not determined and hence where the waves are not "monochromatic," will always satisfy the equation

$$\Delta\psi - \frac{8\pi^2 m}{h^2} U\psi = -\frac{4\pi i m}{h}\frac{\partial\psi}{\partial t}, \tag{136}$$

which we shall call the *time-dependent Schrödinger equation.*

It is to be noted that since there is an imaginary coefficient in (136), the complex conjugate ψ^* of ψ does not satisfy the same equation but does satisfy the following one:

$$\Delta\psi^* - \frac{8\pi^2 m}{h^2} U\psi^* = +\frac{4\pi i m}{h}\frac{\partial\psi^*}{\partial t}. \tag{136'}$$

31. Current density. In analogy to the intensity of illumination, defined statistically in §19, it is convenient to define the (probabilistic) particle flux density. We shall show that there exists a vector **i** such that $i_n\, d\sigma\, dt$ (where $d\sigma$ is an element of surface and n is the normal to it, with a given positive direction) represents the probability, taken algebraically,[18] that the particle crosses the element $d\sigma$ in the time dt.

For this demonstration we shall suppose (as in the footnote on page 145) that there is not one, but a great number N of systems in the same conditions, and not interacting. The mean density of particles will be $|\psi|^2$. Therefore in any enclosed space S there will be

$$N\int_S \psi\psi^*\, dS$$

particles on the average.

The increase of this number in unit time is

$$N\frac{\partial}{\partial t}\int_S \psi\psi^*\, dS = N\int_S\left(\psi\frac{\partial\psi^*}{\partial t} + \psi^*\frac{\partial\psi}{\partial t}\right) dS$$

---

[18] That is, the difference between the probability of crossing in the positive sense and that in the negative sense.

or, obtaining the derivatives from (136) and (136′)

$$N \frac{h}{4\pi i m} \int_S (\psi \, \Delta\psi^* - \psi^* \, \Delta\psi) \, dS.$$

On the other hand, this quantity must be equal to the average number of particles which enter the volume S in unit time across the surface σ. Therefore we must have

$$N \frac{h}{4\pi i m} \int_S (\psi \, \Delta\psi^* - \psi^* \, \Delta\psi) \, dS = N \int_\sigma i_n \, d\sigma$$

(taking the normal to be directed inward). Transforming the volume integral into a surface integral, we have

$$N \frac{h}{4\pi i m} \int_\sigma \left(\psi^* \frac{\partial \psi}{\partial n} - \psi \frac{\partial \psi^*}{\partial n} \right) d\sigma = N \int_\sigma i_n \, d\sigma.$$

In order for this equation to hold for any σ, it suffices to take

$$\mathbf{i} = \frac{h}{4\pi i m} (\psi^* \operatorname{grad} \psi - \psi \operatorname{grad} \psi^*). \tag{137}$$

This is the desired expression for the *probability current density.*

It is to be noted that if the particle is in a "stationary state" (see §29), that is, if ψ has the form (128), the current density becomes

$$\mathbf{i} = \frac{h}{4\pi i m} (u^* \operatorname{grad} u - u \operatorname{grad} u^*) \tag{137′}$$

and is independent of time, just as is the average density $\psi\psi^* = uu^*$. For this reason these states are called *stationary states.*

If we are dealing with particles carrying an electric charge e, the vector $\mathbf{j} = e\mathbf{i}$ obviously represents the average value of the electric current density. ·

In the first works on wave mechanics $\psi\psi^*$ was not interpreted as a probability density or an average density but as a true density, so that the electron was thought of as a continuous distribution of electricity, of electrical charge density $\rho = e\psi\psi^*$. Consequently, the vector $\mathbf{j} = e\mathbf{i}$ defined by (137) was interpreted as a true electric current density (not average), and on this basis the electromagnetic field produced and the emitted radiation were calculated by using the ordinary laws of electromagnetic theory. This hypothesis, although leading to correct results in some cases, was in manifest contrast with all phenomena in which the electron exhibits its corpuscular nature, and was therefore soon abandoned. Born was the first to suggest the probabilistic interpretation, which was later perfectly incorporated into the framework of the uncertainty principle, as has been shown in the preceding sections.

32. Electromagnetic field and radiation. When the particle under consideration is electrically charged (for example, if it is an electron), the problem arises of determining its electric and magnetic effects upon other particles and, in particular, its radiation. This latter question has been answered in a precise manner by a theory developed by Dirac, for which the reader is referred to other works.[19]

We shall limit ourselves here to mentioning that we may get an estimate of the *average* value of the electric field or of the magnetic field (see footnote on page 145) by applying the ordinary laws of electromagnetism and taking for charge and current densities the *average* values

$$\rho = e\psi\psi^*, \tag{138}$$

$$\mathbf{j} = e\mathrm{i} = \frac{he}{4\pi im}(\psi^* \operatorname{grad} \psi - \psi \operatorname{grad} \psi^*). \tag{139}$$

In the case of a stationary state these expressions become

$$\rho = euu^*, \tag{140}$$

$$\mathbf{j} = \frac{he}{4\pi im}(u^* \operatorname{grad} u - u \operatorname{grad} u^*), \tag{141}$$

and are independent of t. In this case, then, the average distribution of charges and currents is *stationary*. Of course, the average values of the electric and magnetic fields will then also be independent of time. In particular, if u is real, we have $\mathbf{j} = 0$; that is, the average distribution of charges is not only stationary but also *static*. Such a solution always exists, the coefficients of (131′) being real; as a matter of fact, in cases where there is degeneracy, it is the only solution (except for an unimportant factor $e^{i\theta}$ with θ constant).

As for the radiation emitted, we shall confine ourselves here to stating the essential results of the Dirac theory. If, at the time $t = 0$, the system finds itself in a stationary state of energy E_n (which is not to be the lowest one) and we observe it at the time t, there is a certain probability, increasing with t, of finding it in a

---

[19] See also E. Fermi, "Quantum Theory of Radiation," *Rev. Mod. Phys.* **4**, pp. 87–132 (1932), or Nos. 15 and 29a of the Bibliography.

state of energy $E_m < E_n$; the difference in energy being emitted in the form of radiation of frequency

$$\nu_{nm} = \frac{E_n - E_m}{h}. \tag{142}$$

(This formula, which in the Bohr theory constitutes a postulate in itself, is derived instead, in the Dirac theory, from the general principles of quantum mechanics). This probability, in every infinitesimal interval dt, increases by $P_{nm}\,dt$, where

$$P_{nm} = \frac{32\pi^4}{3hc^3}\,\nu_{nm}^3(X_{nm}^2 + Y_{nm}^2 + Z_{nm}^2). \tag{143}$$

In this formula, X_{nm}, Y_{nm}, and Z_{nm} represent three integrals into which the eigenfunctions of the two stationary states enter, that is, the initial and the final eigenfunctions;

$$X_{nm} = e\int_S xu_n^* u_m\,dS = X_{mn}^* \tag{144}$$

(and analogous expressions for the other two). If initially we have a great number of atoms in the nth state, the intensity of the radiation of frequency ν_{nm} emitted at every instant will naturally be proportional to their number and to P_{nm}. Hence the latter is essentially determined by the three quantities X_{nm}, Y_{nm}, and Z_{nm}. If, in particular, all three come out to be zero, the transition from the nth to the mth state has zero probability, or is "forbidden"; the corresponding spectral line will be missing. In this manner justification is given to the so-called "selection rules," which were long known by spectroscopists and which were explained in the Bohr-Sommerfeld theory by means of the *correspondence principle* (see §64).

Furthermore, we find that each of the three quantities X_{nm}, Y_{nm}, and Z_{nm} determines a component of the amplitude of the electric field in the emitted light, so that we may also get the state of polarization of the latter from them. These quantities correspond to what in classical theory are the amplitude components of the *electric dipole moment* of the emitting system.[20] If, for instance,

---

[20] It will be seen in §33 of Part III that X_{nm}, Y_{nm}, and Z_{nm} are the *matrix* elements which represent the components of the electric dipole moment in quantum mechanics.

$X_{nm} = Y_{nm} = 0$, $Z_{nm} \neq 0$, the emitted radiation will correspond to that of a linear oscillator parallel to the z-axis, and hence its plane of vibration (plane containing the ray and the electric vector) is parallel to the z-axis.

Analogous arguments may be made in connection with absorption (see also §43 of Part III).

CHAPTER 7

One-dimensional Problems

33. Characteristics of one-dimensional problems. In order to study the Schrödinger equation in the simplest cases, we shall treat in this chapter some "one-dimensional" problems, by which we understand that in U, in ψ, and in all other quantities which eventually enter, only one of the space coordinates (in addition to t) occurs, for instance only x.

The physical characteristic of these problems is that the particle may be found with equal probability at all points of any plane normal to the x-axis, or that its coordinates y and z are completely indeterminate, able to take on any value with equal probability. We are looking only for the probability distribution of x. We can also say that we are studying the motion, not of the particle P, but only of its projection P' on the x-axis. Therefore these problems correspond to those of classical mechanics in which the motion of a point along a straight line is studied. We shall make use of the same terminology and shall speak of the motion of the particle along the x-axis (by which we refer to its projection P'). Incidentally, we note that the particle will necessarily be subject to the conditions which we are now considering, whenever we suppose that the particle is constrained to move *parallel* to the x-axis (not *on* the x-axis). In fact, by virtue of the uncertainty principle, this condition ($p_y = p_z = 0$) implies the complete indeterminacy of the coordinates y and z.

With this premise, (131′) becomes, in our case, an equation with ordinary derivatives:

$$\frac{d^2u}{dx^2} + \frac{8\pi^2 m}{h^2}[E - U(x)]u = 0, \tag{145}$$

which we shall call the *one-dimensional Schrödinger equation* (for stationary states).

We shall now show the application of this equation to some particular problems, with the idea in mind of illustrating more effectively, by means of analytically simple examples, the spirit of wave mechanics; some of these problems are also susceptible of direct physical application.

It will be useful to keep in mind, as an intuitive guide to the solution of these problems, their formal analogy with optical problems in which the light propagates by plane waves along the x-axis and in which the index of refraction (independent of y and z) varies along x according to $\sqrt{E - U}$. In these one-dimensional problems we may also invoke the analogy with the transverse oscillations of a string, making u correspond to the displacement of the string, and its density to $(E - U)$.

34. Qualitative discussion of one-dimensional problems. Equation (145) is of the type

$$\frac{d^2u}{dx^2} + f(x)u = 0, \tag{146}$$

with $f(x) = \frac{8\pi^2m}{h^2}(E - U)$. If we represent the potential U as a function of x by a graph (Fig. 24) and if we draw a horizontal line of ordinate E, the intersections yield the values $x_1, x_2, \ldots$ of x for which $f(x) = 0$. These points divide the x-axis into regions, in some of which $f > 0$, in others $f < 0$. In classical mechanics the first regions (those for which the difference $E - U$, which represents the kinetic energy, is positive) are the only ones in which the particle may move, the total value of the energy E being assigned. On the

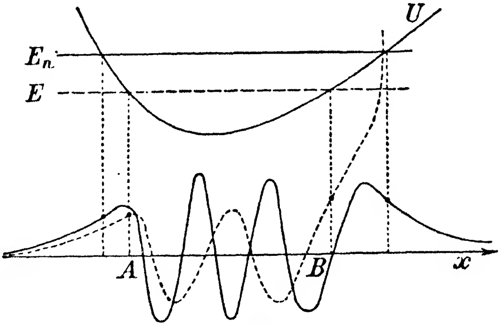

Fig. 24

contrary, the other regions (in which $E - U < 0$) are inaccessible to a particle of energy E. As we shall see, this difference does not hold true in wave mechanics, although even here the distinction of the two regions has an important significance, because the behavior of $u(x)$ is considerably different in them.

We observe first of all that in the regions where $f > 0$, u and $\frac{d^2u}{dx^2}$ have opposite signs, and hence the curve representing u is always *concave* toward the x-axis there; conversely, it is *convex* toward the x-axis in regions in which $f < 0$. It is clear, then, that in the first regions the curve may cross the x-axis several times with an oscillatory behavior; whereas in the second regions the curve, if it crosses the x-axis even once, cannot turn back and cross the axis again (in the same region) since it has to stay convex, and hence one of these regions may at the most contain a single node. Thus the curve is no longer oscillatory in character. This concept becomes clearer if we think of the particular case $f =$ constant. If this constant is positive, the curve is sinusoidal (oscillatory in character, concave toward the x-axis); if it is negative, the curve is of the exponential type (nonoscillatory in character, convex toward the x-axis).

This consideration turns out to be useful in a discussion of the qualitative behavior of the eigenfunctions corresponding to a given potential curve and to a given energy level. For instance, let us consider a potential of the type shown in Fig. 24, that is, a potential possessing a single minimum and tending monotonically toward infinity to the right and left of it in any manner. The harmonic oscillator, for example, belongs to this class (see §39). For any value of E, let us consider a curve u (dashed curve) representing a solution which to the left approaches the x-axis asymptotically. Toward the right it moves away from the x-axis, being convex toward it, up to point A, where it will start a number of oscillations over the portion AB, only to assume, to the right of B, a nonoscillatory behavior which in general will make it move away indefinitely from the x-axis. Only if E has one of the values E_n (eigenvalues), will the curve approach the x-axis to the right too; it will then represent an eigenfunction u_n (solid curve).

The portion over which the curve shows an oscillatory character is evidently the region in which the particle would oscillate according to classical mechanics.

35. Particle of definite energy, not subject to forces. Let us apply the one-dimensional Schrödinger equation to the case of a particle without forces, free to move from $x = -\infty$ to $x = +\infty$.

We may set $U = 0$ in (145), and hence, substituting for convenience

$$k = \sqrt{\frac{2mE}{h^2}}, \tag{147}$$

we may write the equation

$$\frac{d^2u}{dx^2} + 4\pi^2k^2u = 0, \tag{148}$$

k being constant. This is the equation studied in §8 and has for general solution

$$u = a\,e^{2\pi ikx} + b\,e^{-2\pi ikx}, \tag{149}$$

whence

$$\psi = (a\,e^{2\pi ikx} + b\,e^{-2\pi ikx})\,e^{-\frac{2\pi i}{h}Et}. \tag{150}$$

We immediately recognize that E may not be negative (and hence that k must be real) because otherwise u would become infinite for $x \rightarrow +\infty$ or $x \rightarrow -\infty$, which case must be excluded (see §28). Aside from this consideration, k^2 and hence E are completely arbitrary, as we have seen, between 0 and $+\infty$ (continuous eigenvalue spectrum); that is, the particle may have any (positive) energy. Let us suppose E to be fixed once and for all. To this fixed value there correspond two linearly independent solutions represented by the two terms in (149), and thus we are dealing with one of the cases of degeneracy studied in §6.

Let us first consider only the first term; that is, let us set $b = 0$, and take

$$\psi = a\,e^{2\pi i(kx-\nu t)}. \tag{151}$$

This formula represents a *traveling* plane wave train (see §12) of wavelength

$$\lambda = \frac{1}{k} \tag{152}$$

and of (phase) velocity

$$V = \frac{\nu}{k} = \frac{E}{hk}. \tag{153}$$

It is seen that according to classical mechanics, there correspond to the energy E a velocity and a momentum of the particle, given respectively by

$$v = \pm\sqrt{\frac{2E}{m}}, \qquad p = \pm\sqrt{2mE} = \pm hk.$$

Hence for a particle in progressive motion we have

$$p = hk,$$

and then (152) may be identified with the already known formula $\lambda = h/p$, while (151) takes on the form

$$\psi = a\, e^{\frac{2\pi i}{h}\left(px - \frac{p^2}{2m}t\right)}. \tag{151'}$$

The constant a is determined by the normalization procedure for continuous spectra explained in §10. We note that the probability density for the position of the particle, given by $|\psi|^2$, turns out to be independent of x. This result is in agreement with the fact that, since the momentum is determined with infinite accuracy, the position of the particle remains completely uncertain, in accord with the uncertainty principle.

Analogous considerations may be made concerning the second term of (150), which we may also put in the form (151′) setting $p = -hk$. This term then represents a wave traveling in the opposite direction and corresponds to a particle whose momentum is directed in the negative sense. We see that the degeneracy of the problem in wave mechanics has its origin in the fact that in ordinary mechanics there correspond two values of p to a given value of E, one positive and one negative.

Considering now the general solution (150) and keeping the principle of superposition in mind, we can say that this solution represents the case in which the energy of the particle is determined while the direction of its momentum is not determined, so that there is a certain probability, proportional to $|a|^2$, of finding the particle in forward motion, and a probability, proportional to $|b|^2$, of finding it in backward motion.

36. Particle of undetermined energy (wave group). In practice, the most interesting case is the more general one in which the probability density $P_0(x)$ for the abscissa of the particle is prescribed at time $t = 0$, as well as the probability density $Q_0(p)$ of the initial momentum. These two functions will be represented by curves whose shapes depend on the type of experiment used to define the initial state of the system.

ψ will generally be of the form (133′), as has been pointed out in §29 (sum or integral of an infinite number of stationary solutions corresponding to the different values of E). Since, however, as has been shown above, to each value of E there correspond two solutions differing in the sign of p, it is convenient to characterize the individual components by the value of p rather than by the value of E, or else to write

$$\psi = a \int_{-\infty}^{\infty} \varphi_0(p)\, e^{\frac{2\pi i}{h}\left(px - \frac{p^2}{2m}t\right)}\, dp, \tag{154}$$

where a is a normalization constant which will be fixed later and $\varphi_0(p)$ has a meaning analogous to $c(E)$ in (133′), that is,

$$|\varphi_0(p)|^2 = Q_0(p). \tag{155}$$

It is to be noted that this relation determines only the absolute value of $\varphi_0(p)$, leaving the argument θ arbitrary. Hence we shall write

$$\varphi_0(p) = \sqrt{Q_0(p)}\, e^{i\theta(p)}, \tag{155′}$$

and the function $\theta(p)$ must be determined in such a way as to satisfy the other initial condition, that is, so that

$$|\psi_0|^2 = P_0(x), \tag{156}$$

where ψ_0 is ψ for $t = 0$:

$$\psi_0(x) = a \int_{-\infty}^{\infty} \varphi_0(p)\, e^{\frac{2\pi i}{h}px}\, dp. \tag{157}$$

We shall see shortly under what conditions this is possible.

It will be observed that (157) may be identified with (58), §12, by identifying ψ_0 with f and setting $p = hk$, and

$$A(k) = a\varphi_0(hk)h; \tag{158}$$

consequently, (59) yields

$$\varphi_0(p) = \frac{1}{ah} \int_{-\infty}^{\infty} \psi_0(x)\, e^{-\frac{2\pi i}{h}px}\, dx. \tag{157′}$$

Noting that by virtue of (51′) of §12, we have

$$\int_{-\infty}^{\infty} \psi_0\psi_0^*\, dx = haa^* \int_{-\infty}^{\infty} \varphi_0\varphi_0^*\, dp$$

and that, because of the normalization condition, both integrals of this formula must be equal to 1, we find that we may put[1]

$$a = \frac{1}{\sqrt{h}} \tag{159}$$

into the preceding formulas.

In the most interesting cases, the initial position of the particle is defined with a certain approximation, so that the curve of $P_0(x)$ is bell-shaped, as is, for example, the Gaussian error curve. We may then say that (154) represents a "wave group" which initially has the "profile" (see §14) defined by the curve $P_0(x)$, and then displaces itself (while deforming) with a group velocity given by (74), which in the present case immediately gives $v = p/m$. The group velocity may thus be identified with the particle velocity in classical mechanics, as is to be expected, by virtue of the manner in which the Schrödinger equation was constructed.

The uncertainty Δx in the abscissa of the particle at a given instant is defined by the formula [analogous to (65)]

$$(\Delta x)^2 = \int (x - \bar{x})^2 \psi\psi^* \, dx \tag{160}$$

and similarly for the uncertainty Δp in momentum

$$(\Delta p)^2 = \int (p - \bar{p})^2 \varphi\varphi^* \, dp, \tag{161}$$

a formula which, when the relation between p and k and (158) is recalled, coincides with (63). Thus we see that the two uncertainties Δx and Δp are subject to the limitation imposed by (66), which is now written

$$\Delta p \, \Delta x \geq \frac{h}{4\pi}; \tag{162}$$

hence the two curves $P_0(x)$ and $Q_0(x)$ may not be assigned arbitrarily. If they are assigned in a manner such that this condition is not satisfied initially, there exists no $\theta(p)$ which when put into (155′) satisfies (156); therefore it is not possible to construct a wave group having the required properties. Thus the scheme of wave mechanics is verified, with respect to the uncertainty principle, for material particles. This result was to be expected, since the formalism of wave mechanics is analogous to that of wave optics.

---

[1] The arbitrary constant of modulus 1 by which a could be multiplied, and hence also ψ, has no influence upon the probability, and hence is of no importance.

Of particular interest is the case in which the curves $P_0(x)$ and $Q_0(x)$ are such that at time zero we have

$$\Delta x \, \Delta p = \frac{h}{4\pi} \tag{163}$$

(that is, such that the product of the uncertainties is the least possible). As we have seen in §13, this case requires that the functions $\psi_0(x)$ and $\varphi_0(p)$ be of the form (70) and (72). In the present notation, if for brevity we set

$$\alpha = \frac{1}{\sqrt{2}(\Delta x)_0}, \qquad \beta = \frac{1}{\sqrt{2}(\Delta p)_0} \tag{164}$$

(that is, if we indicate by α and β the "precisions"[2] of the initial determinations of x and p), the functions $\psi_0(x)$ and $\varphi_0(p)$ become

$$\psi_0(x) = C\, e^{-\frac{\alpha^2}{2}(x-\bar{x}_0)^2+\frac{2\pi i}{h}\bar{p}_0 x}, \tag{165}$$

$$\varphi_0(p) = D\, e^{-\frac{\beta^2}{2}(p-\bar{p}_0)^2-\frac{2\pi i}{h}\bar{x}_0 p}. \tag{166}$$

The curve of the initial probabilities are thus of the Gaussian type, as follows:

$$P_0(x) = CC^*\, e^{-\alpha^2(x-\bar{x}_0)^2}, \tag{167}$$

$$Q_0(x) = DD^*\, e^{-\beta^2(p-\bar{p}_0)^2}. \tag{168}$$

Equation (163) is equivalent to the following relation between α and β

$$\alpha\beta = \frac{2\pi}{h}. \tag{163'}$$

The moduli of the constants C and D are determined by the normalization conditions, which yield

$$CC^* = \frac{\alpha}{\sqrt{\pi}}, \tag{169}$$

$$DD^* = \frac{\beta}{\sqrt{\pi}}. \tag{170}$$

Let us now calculate the probability curve $P(x)$ of position at the time t.

---

[2] Note that the terminology of error theory is being applied to a type of uncertainty of an entirely different origin from that of the errors of observation.

We must first of all calculate ψ at time t by means of (154). If we introduce expression (166) and set for short

$$x - \bar{x}_0 = \xi, \qquad p - \bar{p}_0 = \eta,$$

(154) becomes

$$\psi = \frac{D}{\sqrt{h}} e^{\frac{2\pi i}{h}\left[\bar{p}_0\xi - \frac{\bar{p}_0^2}{2m}t\right]} \int_{-\infty}^{\infty} e^{-\left(\frac{\beta^2}{2}+\frac{\pi i t}{hm}\right)\eta^2 + \frac{2\pi i}{h}\eta\left(\xi - \frac{\bar{p}_0}{m}t\right)}\, d\eta.$$

The integral may be easily reduced to known definite integrals.[3] It proves to be

$$\frac{\sqrt{\pi}}{\sqrt{\frac{\beta^2}{2} + \frac{\pi i t}{hm}}} e^{-\frac{\pi^2}{h^2}\frac{\left(\xi - \frac{\bar{p}_0}{m}t\right)^2}{\frac{\beta^2}{2}+\frac{\pi i t}{hm}}}. \tag{171}$$

After ψ has been calculated thus, $\psi\psi^*$ will give the function $P(x)$ at time t. In order to write the square of the modulus of expression (171) in a simple form, it is convenient to introduce the notation

$$\alpha_t = \frac{\pi\beta}{h} \frac{1}{\sqrt{\frac{\beta^4}{4} + \left(\frac{\pi t}{hm}\right)^2}}. \tag{172}$$

Thus we obtain

$$P(x) = DD^* \frac{\alpha_t}{\beta} e^{-\alpha_t^2\left(\xi - \frac{\bar{p}_0}{m}t\right)^2};$$

or, if we note that because of (169) and (170)

$$\frac{DD^*}{\beta} = \frac{CC^*}{\alpha}$$

and introduce x again:

$$P(x) = CC^* \frac{\alpha_t}{\alpha} e^{-\alpha_t^2\left(x - \bar{x}_0 - \frac{\bar{p}_0}{m}t\right)^2}. \tag{173}$$

This formula, when compared with (167), shows that the probability curve at time t is still a Gaussian, but its maximum, instead of being at $x = \bar{x}_0$, is at

$$x = \bar{x}_0 + \frac{\bar{p}_0}{m} t;$$

---

[3] Specifically, to the integral $\int_{-\infty}^{\infty} e^{-a^2x^2 - bx}\, dx = \frac{\sqrt{\pi}}{a} e^{\frac{b^2}{4a^2}}$.

that is, the maximum moves with velocity $\bar{p}_0/m$, as we already know. Furthermore, the precision is no longer α but α_t, given by (172), which on account of (163′) may also be written

$$\alpha_t = \frac{\alpha}{\sqrt{1 + \left(\frac{\alpha t}{\beta m}\right)^2}}. \tag{172′}$$

This formula shows that the precision diminishes with time, that is, that the wave group spreads out more and more as it propagates (after time $t = 0$), in correspondence with the fact (also valid in ordinary mechanics) that when the initial velocity is not exactly determined, the uncertainty in position grows with time. We note that if we want to make use of a measurement made at time zero in order to calculate the position at a *previous* instant ($t < 0$), again we find an indeterminacy which is the larger, the further that instant is removed from the instant of measurement.

37. Potential step. Let us suppose that a particle of definite energy E is subject to a potential $U(x)$ represented in Fig. 25.

$$\begin{aligned} U &= 0 \qquad \text{for } x < 0 \quad \text{(region I)}, \\ U &= U_0 \qquad \text{for } x > 0 \text{ (region II)}. \end{aligned}$$

This purely idealized condition may be considered as the limiting case of the following. Let us suppose that over a segment PP' there exists a force field which forces the particle toward the left (as could be realized, in the case of an electron, by two charged grids of opposite sign located at P and P'). Then the potential would have the shape $APQC$. Now letting the interval PP' tend to zero while maintaining the difference of potential between P and P' (and hence making the field become infinite), we reach the limiting case in question. It may therefore be considered a "double-layer potential."

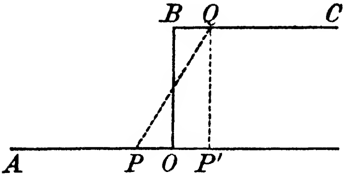

Fig. 25

The Schrödinger equation in region I will still be given by (148), whereas in region II it will have the same form except that instead of k there enters the constant

$$k_0 = \sqrt{\frac{2m}{h^2}(E - U_0)},$$

which may be either real or imaginary, according to whether $E \geq U_0$ or $E < U_0$.

The function u will therefore be given by

$$u = A_1 e^{2\pi ikx} + B_1 e^{-2\pi ikx} \quad (\text{for } x < 0), \qquad (174)$$

$$u = A_2 e^{2\pi ik_0x} + B_2 e^{-2\pi ik_0x} \quad (\text{for } x > 0); \qquad (174')$$

and since u must be continuous, together with its first derivative, for $x = 0$, the four constants A_1, A_2, B_1, B_2 must be connected by the relations

$$\left.\begin{aligned} A_1 + B_1 &= A_2 + B_2 \\ A_1 - B_1 &= \mu(A_2 - B_2) \end{aligned}\right\}, \qquad (175)$$

where we have put $\mu = \frac{k_0}{k} = \sqrt{1 - \frac{U_0}{E}}$.

It is now convenient to distinguish two cases according to whether the energy E of the particle is (a) above or (b) below the difference U_0 of the potential levels. We shall suppose in each case that the particle is moving from $-\infty$ toward O.

Case (a): $E \geq U_0$. In this case, according to classical mechanics, the particle would overcome the potential step and would continue its motion to the right of O with a velocity reduced in the ratio $1:\mu$. From the point of view of wave mechanics, however, we note that each of relations (174) and (174′) represents the superposition of forward and backward moving waves. Since we suppose that the particle comes from $-\infty$ and not from $+\infty$, there must not be any waves moving backward in region II, and hence $B_2 = 0$. Equations (175) then yield, upon elimination of A_2,

$$B_1 = \frac{1 - \mu}{1 + \mu} A_1.$$

In general, then, there will be in region I, in addition to forward-moving waves of amplitude A_1, waves moving in the opposite direction, *reflected* from the potential step, with amplitude B_1. The meaning of these constants can be stated more easily if we think not of a single moving particle but of a large number of particles. Then $A_1A_1^*$ is proportional to the number of particles (per unit length) which are moving forward in the region to the left of O, whereas $B_1B_1^*$ is proportional to the number of particles moving in the opposite direction. Therefore the ratio

$$R = \frac{B_1 B_1^*}{A_1 A_1^*} = \left(\frac{1 - \mu}{1 + \mu}\right)^2$$

represents the fraction of particles which are reflected, or the probability that one particle will be reflected. Therefore R may be called the *reflection coefficient* of the step, and

$$\tau = 1 - R$$

represents the *transmission coefficient.*

Hence an incident particle has, according to wave mechanics, a certain probability R of being reflected (with the same velocity) and a certain probability τ passing over the step (with a velocity reduced in the ratio $1 : \mu$, since the wavelength $1/k$ becomes $1/k_0$).

The problem is analogous to the reflection of light from a semi-reflecting surface; indeed, the reflection coefficient in optics represents the probability of an incident photon being reflected.

Case (b): $E < U_0$. In this case, according to classical mechanics, the particle would be thrown back without ever passing over the step. In the wave treatment, instead, we must observe that k_0 and μ are imaginary. Let us therefore put

$$k_0 = ik', \qquad \mu = i\mu',$$

with k' and μ' real, and let us write (174′) in the form

$$u = A_2\, e^{-2\pi k' x} + B_2\, e^{2\pi k' x} \qquad (x > 0). \quad (176)$$

We then note that for u to remain finite for $x \to \infty$ we must have $B_2 = 0$. If this stipulation is kept in mind, (175) immediately gives

$$B_1 = \frac{1 - i\mu'}{1 + i\mu'} A_1.$$

Since $1 + i\mu'$ and $1 - i\mu'$ have the same modulus, we can see that the reflected waves have an amplitude equal to the incident waves; that is, we have

$$R = \frac{B_1 B_1^*}{A_1 A_1^*} = 1,$$

which means that all the particles are reflected. However, u is different from zero even to the right of O, where it is given by (176), which reduces to

$$u = A_2 e^{-2\pi k' x}. \tag{177}$$

Hence the real part of u (and similarly the imaginary part) has the shape shown in Fig. 26. Consequently, if we perform an observation of position of the particle, there is a certain probability of finding it also to the right of O, a probability which is appreciable up to a distance of the order of $1/k'$ from O. In this result wave mechanics differs sharply from classical mechanics.

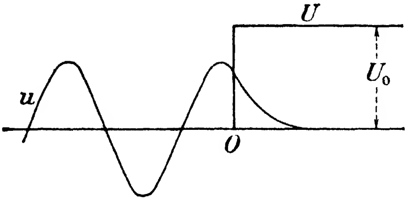

Fig. 26

This result might appear paradoxical from the classical point of view because in the region to the right of O, the potential energy U_0 of the particle would be greater than its total energy E. Hence the kinetic energy $E - U_0$ would be negative and the velocity imaginary. This is the reason why this region is inaccessible to the particle according to classical mechanics. However, from the point of view of wave mechanics, the result is not paradoxical at all. In fact, we must not think that the result means that the particle penetrates a little into the potential step before it is reflected (an interpretation contrary to the spirit of wave mechanics, according to which the motion of a particle may not be followed); instead, it means that upon making an observation of position, we may find the particle also to the right of O. Now it must be remembered that a direct observation to localize the particle inevitably requires that a (unknown) momentum be imparted to the particle, and hence that its energy will be altered. Therefore, upon finding the particle to the right of O, we cannot say that we have found a violation of the principle of conservation of energy, since the fact may be interpreted, in corpuscular terminology, by saying that the energy necessary to overcome the step was imparted to the particle in the very act of observation.[4]

[4] This consideration may be rendered more precise by means of a more detailed analysis which we shall not make here. We restrict ourselves here to pointing out briefly the following objection which apparently may be made to the argument above: Is it not possible to perform the observation in such a way as to have an uncertainty Δp in momentum less than the momentum which classically is necessary to overcome the step (that is, less than

The phenomenon here considered is analogous to a known phenomenon of optics, where in the total reflection of light there also exists a luminous disturbance in the second medium, in the immediate neighborhood of the reflecting surface. However, in the case of light waves total reflection occurs only if the angle of incidence is greater than the critical angle, whereas for de Broglie waves the phenomenon occurs even at normal incidence, as we have seen.

38. Particle on a limited line segment. We shall now show a first example of quantization by the Schrödinger method, considering a particle which may move (without forces) on a *limited* portion of the line AB, recoiling elastically at the endpoints. We shall look first of all for the stationary solutions (that is, with E definitely determined).

The Schrödinger equation will still be given by (148), but with the condition that ψ vanish outside of the segment AB, the probability of finding the particle there being zero by hypothesis.

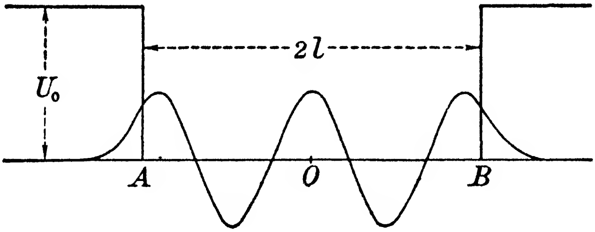

Fig. 27

The conditions of this idealized problem may be expressed in a manner which is more apt to show its physical significance if we will think of the two obstacles A and B as being two *potential steps* of height U_0 (Fig. 27) and then imagine that U_0 tends toward infinity.

$\sqrt{2m}(\sqrt{U_0} - \sqrt{E})$), while being resigned to having a $\Delta x \geq \frac{h}{4\pi\,\Delta p}$ by the uncertainty principle? In that case, if from the measurement $x > \Delta x$, in spite of the uncertainty in x we should be certain that the particle is to the right of O. This reasoning is erroneous because Δx and Δp represent the average values of the spreads and not their maximum values which, by a theorem stated in §13, cannot both be finite. Furthermore, since the position of the particle is not a priori altogether uncertain (it being practically certain that $x < 1/k'$), it is not possible to make Δp as small as we wish (in fact, the momentum of the particle even before the observation is $\pm\sqrt{2mE}$, and hence has an uncertainty $2\sqrt{2mE}$).

In this manner k' of (177) approaches ∞, and hence the exponential of Fig. 26 becomes equal to zero immediately beyond the step. We can thus fix our attention only upon the behavior of ψ within the segment AB, imposing the condition that ψ must vanish at the endpoints.

Since equation (148) is identical with the one already studied in §8, with $4\pi^2k^2$ made to correspond to the parameter λ of (21), the present problem has already been treated in §8 under conditions (α). It follows then from (24′) that k may take on only the values

$$k_n = \frac{n}{4l} \qquad n = 1, 2, 3, \ldots .$$

Evidently, these values are those for which the half wavelength $1/2k_n$ is a submultiple of the length $2l$. From this equation we obtain, by means of (147), the energy levels, which are

$$E_n = \frac{h^2}{32ml^2}\,n^2. \tag{178}$$

To these there correspond the (normalized) eigenfunctions given by (25):

$$u_n = \sqrt{\frac{1}{l}}\,\sin n\,\frac{\pi}{2l}\,(x + l), \tag{179}$$

and hence

$$\psi_n = \sqrt{\frac{1}{l}}\,e^{-\frac{2\pi i}{h}E_nt}\,\sin n\,\frac{\pi}{2l}\,(x + l). \tag{179'}$$

The probability density of position at any instant is

$$P(x) = \psi\psi^* = uu^* = \frac{1}{l}\sin^2 n\,\frac{\pi}{2l}\,(x + l) \tag{180}$$

and hence possesses nodes which divide AB into n equal parts.

The probability distribution of the momentum is obtained by noting that (179′) may be written

$$\psi_n = c_1\,e^{i\frac{n\pi}{2l}x - \frac{2\pi i}{h}E_nt} + c_2\,e^{-i\frac{n\pi}{2l}x - \frac{2\pi i}{h}E_nt},$$

where c_1 and c_2 are given by (23′) of §8. Since the first of these two terms represents forward-moving waves of "wave number" k_n and the second term represents identical waves moving in the opposite direction, at every instant there will be a probability proportional to $|c_1|^2$ of finding the particle with momentum $p_n = hk_n$ (see page 169) and a probability proportional to $|c_2|^2$ of finding it with momentum $p_n = -hk_n$. From equations (23′) we then find that c_1 and c_2 have the same modulus, and hence the two probabilities are equal. Therefore the probability of the momentum is not continuously distributed but is zero everywhere except for the two values $\pm hk_n$, where it is equal to ½. It is also verified that the product of the uncertainty in x (that is, $2l$) and the uncertainty in p (that is, $2hk_n$) is nh, and hence of an order of magnitude not less than h.

Let us now proceed to the case in which E does not have a definite energy (or where the system is not in a stationary or quantum state). In this case ψ is represented by a sum of terms of the form (179′);

$$\psi = \sqrt{\frac{1}{l}} \sum_{n=1}^{\infty} c_n \sin n\pi \frac{x + l}{2l} e^{-\frac{2\pi i}{h}E_n t}, \tag{181}$$

and $|c_n|^2$ represents the probability that the system is in the nth state, that is, that its energy is E_n. As we know, the coefficients c_n are subject to the restriction $\Sigma|c_n|^2 = 1$, which is equivalent to the normalization of ψ.

In particular, the c_n may be determined so as to constitute a small wave group, and we find then that this group moves back and forth between A and B in a way similar to the motion predicted by classical mechanics but spreading out gradually, like the group considered in §36.

39. The harmonic oscillator. A *harmonic oscillator* is a system made up of a material point particle moving along a straight line and attracted toward a point O of the line with a force proportional to the distance. If we take the line to be the x-axis and O the origin, the force acting upon the point will thus be $-Kx$, where K is a positive constant. It is known that according to classical mechanics the point will perform simple harmonic oscillations about O with arbitrary amplitude and phase (depending upon the initial conditions), and frequency

$$\nu_0 = \frac{1}{2\pi}\sqrt{\frac{K}{m}}. \tag{182}$$

In order to investigate the corresponding problem in wave mechanics, we note that the potential energy corresponding to the force $-Kx$ is

$$U = \frac{K}{2}x^2 = 2\pi^2 m\nu_0^2 x^2; \tag{182'}$$

and hence, putting this into the one-dimensional Schrödinger equation (146), we obtain

$$\frac{d^2u}{dx^2} + \frac{8\pi^2 m}{h^2}(E - 2\pi^2 m\nu_0^2 x^2)u = 0. \tag{183}$$

We are to find the eigenvalues and eigenfunctions of this equation for the interval $-\infty$ to $+\infty$.

Let us first of all reduce the equation to a form which will put its analytical aspect clearly in evidence. In order to do this, it suffices to change the scale of x by introducing the variable

$$\xi = x\sqrt{\frac{4\pi^2 m\nu_0}{h}}$$

and then by dividing the entire equation by $4\pi^2 m\nu_0/h$, after which it becomes

$$\frac{d^2u}{d\xi^2} + (\epsilon - \xi^2)u = 0, \tag{183'}$$

where we have abbreviated by ϵ the constant

$$\epsilon = \frac{2E}{h\nu_0}, \tag{184}$$

which represents the energy measured in units of $h\nu_0/2$.

Now we have to find whether (183') possesses solutions which are finite and continuous everywhere and which tend to 0 for ξ tending toward $\pm\infty$. It will be found that such solutions are possible only if ϵ is equal to an odd positive integer.

Equation (183') may be discussed in the following manner. First of all we note that its coefficients are finite for all finite values of ξ, hence any possible singularities of u may occur only for $\xi = \pm\infty$. With the criterion of §16 we recognize that in effect the equation has two singularities of the non-Fuchsian type at infinity. An idea of the asymptotic behavior of the solutions at these points may be had by the following heuristic method. If we attempt to satisfy the equation by letting $u = e^{\lambda\xi^2}$, we shall find that, neglecting the terms which are finite with respect to those in ξ^2, the equation may be satisfied asymptotically by taking $\lambda = \pm\frac{1}{2}$. Since we are looking for solutions which do not go to infinity, we must discard those having $e^{\frac{1}{2}\xi^2}$ for asymptotic expression. Let us therefore put

$$u = e^{-\frac{1}{2}\xi^2}v, \tag{185}$$

and let us find whether we can determine v so as to satisfy the equation exactly and so that in addition u tends toward zero at infinity, for which it suffices that v have no essential singularities there.

Substituting (185) into (183′) we find for v the equation

$$v'' - 2\xi v' + (\epsilon - 1)v = 0. \tag{186}$$

Let us try to integrate this equation by a series of the form

$$v = \sum_{r=0}^{\infty} a_r\xi^r, \tag{187}$$

from which we exclude the negative powers of ξ, since we want the solution to be finite also for $\xi = 0$. Substituting this series into (186), we find

$$\sum_{r=0}^{\infty} [(r+1)(r+2)a_{r+2} + (\epsilon - 1 - 2r)a_r]\xi^r = 0.$$

In order for this equation to be satisfied identically, all coefficients must vanish; this condition gives for a_r the recursion relation

$$a_{r+2} = \frac{2r + 1 - \epsilon}{(r+1)(r+2)}\, a_r. \tag{188}$$

It is characteristic of this equation that it relates each coefficient to the one preceding it by *two;* thus, having fixed an arbitrary a_0, we obtain from it by means of (188) all the coefficients with even index; and by fixing an arbitrary value for a_1, all the coefficients with odd index. Two fundamental solutions are obtained, for instance, by taking

$$a_0 \neq 0, \qquad a_1 = 0 \quad \text{(series of even powers)} \tag{189}$$

or else

$$a_0 = 0, \qquad a_1 \neq 0 \quad \text{(series of odd powers),} \tag{189′}$$

and any other solution will be a linear combination of these two.

Now it may easily be shown that, in general, these two series have essential singularities for $\xi = \pm\infty$. Only in the case that one of the coefficients, for example a_{n+2}, vanishes is this not so, because then, according to (188), all the successive coefficients will also vanish and the series reduces to a polynomial of the nth degree. The condition for $a_{n+2} = 0$ with $a_n \neq 0$ is, as may be seen from (188), that

$$\epsilon = 2n + 1. \tag{190}$$

Therefore the eigenvalues of (183′) are all the positive odd integers.

From (184) and (190) it follows that E must have one of the values

$$E_n = (n + \tfrac{1}{2})h\nu_0 \quad (n = 0, 1, 2, \ldots). \tag{191}$$

Hence these are the energy levels of the harmonic oscillator. It may easily be verified that the so-called "zero-point energy" ($E_0 = \frac{1}{2}h\nu_0$) is a necessary consequence of the uncertainty principle.

In this connection we note that the old Bohr-Sommerfeld theory yields, as we shall see in §54, the formula $E_n = nh\nu_0$ instead of (191). (This may be remembered by stating it as follows: the energy of the harmonic oscillator must be an integral multiple of the energy of a photon having a frequency equal to the characteristic frequency of the oscillator.) The difference between two successive energy levels, occurring in the phenomena of emission and absorption, is therefore the same in both theories; but the lowest level, which is zero in the Bohr-Sommerfeld theory, turns out to be $\frac{1}{2}h\nu_0$. A great number of experimental data (band spectra, for example) confirm formula (191).

Let us now proceed to an examination of the eigenfunctions. When ϵ is given by (190), the recursion formula (188) becomes

$$a_{r+2} = -\frac{2(n-r)}{(r+1)(r+2)}\,a_r. \tag{192}$$

Also, if n is even, we must consider the solution in even powers ($a_1 = 0$, $a_0 \neq 0$, and arbitrary); if n is odd, we must take the solution in odd powers ($a_0 = 0$, $a_1 \neq 0$, and arbitrary). Such polynomials occur in the development of the successive derivatives of the function $e^{-\xi^2}$. In fact, it may easily be verified that

$$\frac{d^n}{d\xi^n}\,e^{-\xi^2} = (-1)^n H_n(\xi)\,e^{-\xi^2}, \tag{193}$$

where $H_n(\xi)$ is a polynomial of the nth degree, of the type we are considering. These H_n are called *Hermite polynomials.*[5] The first four of these, for instance, are (as may easily be found)

$$H_0 = 1, \qquad H_1 = 2\xi,$$
$$H_2 = -2 + 4\xi^2, \qquad H_3 = -12\xi + 8\xi^3.$$

Hence to the eigenvalue E_n there corresponds a v given by

$$v_n = \frac{1}{N_n} H_n(\xi)$$

(where N_n is a constant normalization factor) and hence, by (185), a u given by

$$u_n = \frac{1}{N_n} H_n(\xi)\, e^{-\frac{1}{2}\xi^2}.$$

The normalization condition is

$$\int_{-\infty}^{\infty} u_n^2\, dx = \sqrt{\frac{h}{4\pi^2 m\nu_0}} \int_{-\infty}^{\infty} u_n^2\, d\xi = 1$$

or else
$$N_n^2 = \sqrt{\frac{h}{4\pi^2 m\nu_0}} \int_{-\infty}^{\infty} H_n^2(\xi)\, e^{-\xi^2}\, d\xi.$$

The integral is calculated by successive integrations by parts, use being made of the following property of the Hermite polynomials:

$$\frac{dH_n}{d\xi} = 2nH_{n-1}, \tag{194}$$

and we find
$$N_n = \sqrt[4]{\frac{h}{4\pi^2 m\nu_0}}\, \sqrt{2^n n!\sqrt{\pi}}.$$

By a similar procedure we may verify that

$$\int_{-\infty}^{\infty} H_n(\xi) H_m(\xi)\, e^{-\xi^2}\, d\xi = 0$$

for $n \neq m$, that is, that the functions u_n are orthogonal.

The functions u_n for the first five values of n are represented in Fig. 28, neglecting the factor $\sqrt{h/(4\pi^2 m\nu_0)}$ (the first of these functions, $n = 0$, is just the well-known Gaussian). The function u_n^2, which gives the probability distribution of position of the particle if it is known that the latter is in the nth quantum state, is repre-

---

[5] See, for instance, No. 25 or No. 34 of the Bibliography.

sented in Fig. 29 for the first five values of n. In these diagrams the segment OC represents the amplitude of the oscillations according to the classical model ($OC = \sqrt{2E_n/K}$). In this connection we note that for $x > OC$ we have $U > E_n$, and hence the kinetic energy ($E_n - U$) for a particle to the right of C turns out to be negative, or its velocity imaginary. Therefore that region is inaccessible to the particle of energy E_n, according to classical mechanics. Nevertheless, as is shown by the curves of Fig. 29, there is the possibility of finding the particle in that region. This is an apparent paradox analogous to the one already explained in §37.

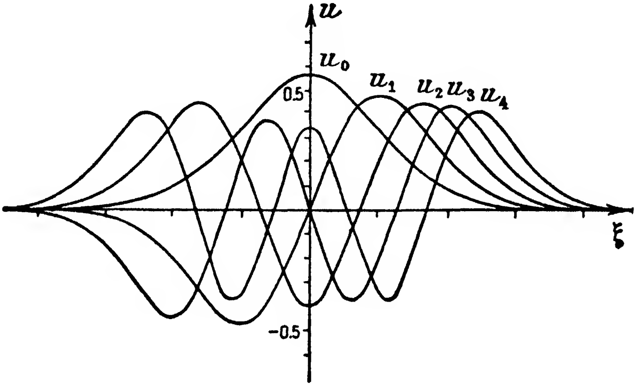

Fig. 28

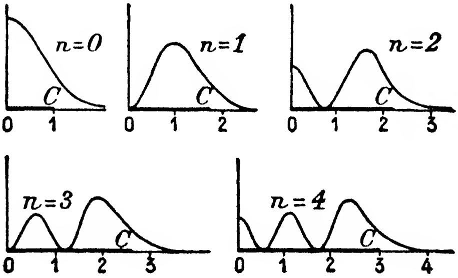

Fig. 29

We now recall that if, instead of the energy of the system being determined, the initial position and velocity of the particle are ascertained (with the approximations required by the uncertainty principle), then the probability distribution, for a consequent determination of the x coordinate, is given by the square of the modulus of a ψ, which is obtained as a linear combination of solutions corresponding to various quantum states. This combination represents a wave group which is the more limited the more precisely the initial determination of x was made. Schrödinger has shown that this wave group will move with an oscillatory motion approximating the harmonic motion of classical mechanics. In this case, however, in contrast to what is encountered in general (see §36, for example), the wave group will retain its initial dimensions indefinitely without spreading out.

40. Potential barrier. Of interest for its applications (some of which will be mentioned in what follows) is the problem of the motion of a particle along a line x, under the action of a potential which is zero everywhere except over a certain interval AB, where the potential increases to a maximum U_0 and then falls to zero again; this behavior is shown qualitatively in Fig. 30. The region AB constitutes what is called a *potential barrier*. A particle which, for instance, originates to the left feels no force up to A, where it encounters a retarding force from A to C', and from there on an accelerating force from C' to B, then again no force at all.

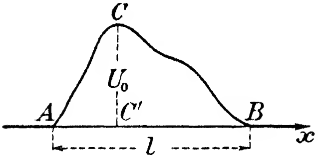

Fig. 30

According to classical mechanics, the particle will overcome the barrier if its initial kinetic energy E is greater than the maximum U_0 of the potential; otherwise it will be thrown backward before reaching C'.

Wave mechanics instead gives a different result: the incident particle has in every case a certain probability of passing over the barrier (even if $E < U_0$) and a certain probability of being repelled (even if $E > U_0$). In order to find this result in a simple manner, we shall idealize the problem again by giving to the potential the shape shown in Fig. 31;

$$\begin{aligned} U &= 0 \quad \text{for } x < 0 \text{ and } x > l \quad \text{(regions I and III)}, \\ U &= U_0 \quad \text{for } 0 < x < l \quad \text{(region II)}. \end{aligned}$$

In the more general case of Fig. 30, a process of successive approximations would give a result qualitatively analogous to the one which we shall find by giving the barrier the particular form of Fig. 31, made up of two "potential steps" in opposite directions.

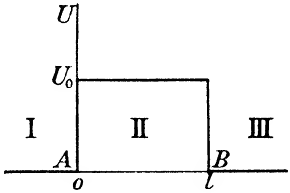

Fig. 31

Hence let us consider the three regions (I, II, III) separately. For regions I and III, the Schrödinger equation is the same as (148), which was studied in §35. Its general solution is of the form (149), but the constants occurring in it will in general be different in the two sections. In region II, the solution will have the same form except for the substitution of $E - U_0$ in the place of E. We can therefore write

$$\left.\begin{aligned} \text{Region I: } u &= A_1 e^{2\pi ikx} + B_1 e^{-2\pi ikx}, \\ \text{Region II: } u &= A_2 e^{2\pi ik_0x} + B_2 e^{-2\pi ik_0x}, \\ \text{Region III: } u &= A_3 e^{2\pi ikx} + B_3 e^{-2\pi ikx}, \end{aligned}\right\} \tag{195}$$

where we have put

$$k = \sqrt{\frac{2m}{h^2} E}, \qquad k_0 = \sqrt{\frac{2m}{h^2} (E - U_0)}.$$

Let us suppose that a large number of particles are projected against the barrier from left to right. Then the squares of the moduli of the constants A and B will have the meaning already illustrated in §37. We must also place here $B_3 = 0$ to express the fact that no particles come from the right. The particles will evidently have the same velocity in regions I and III, and the *transmission coefficient* of the barrier (that is, the probability each incident particle has of overcoming it) will be given by

$$\tau = \frac{A_3A_3^*}{A_1A_1^*}. \tag{196}$$

We note that the five constants A_1, B_1, A_2, B_2, and A_3 must be related by the condition that u is to be continuous, along with its derivative, at the points A and B. The real part of u (as well as its imaginary part) will be represented in regions I and III by two sine curves of wavelength $\lambda = 1/k$ (since k is certainly real).

In region II, however, the curve may have two different shapes according to whether the kinetic energy is above or below U_0. In the first case k_0 is real, and hence the curve is sinusoidal also in region II but with a wavelength $\lambda_0 = 1/k_0$ greater than in I and III (Fig. 32). In the second case k_0 is imaginary, and hence the curve in region II is of the exponential type (Fig. 33). In both of these cases the curves of the three regions join continuously, as shown in the figures. This implies a relation between the amplitudes of the two outer sine curves, whose analytic expression we shall find, leading to the calculation of τ.

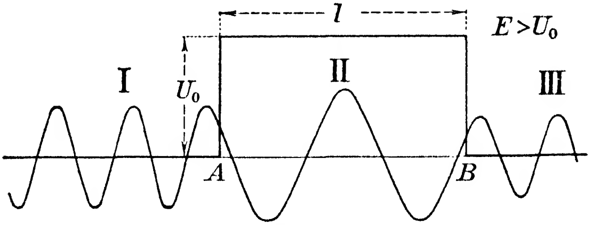

Fig. 32

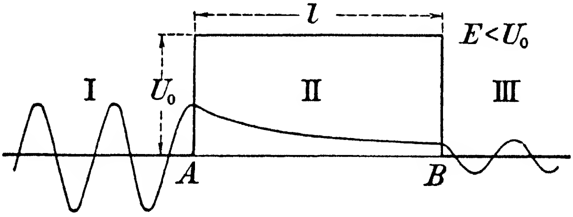

Fig. 33

The continuity conditions on u and $\frac{du}{dx}$ for $x = 0$ give

$$\begin{aligned} A_1 + B_1 &= A_2 + B_2, \\ A_1 - B_1 &= \mu(A_2 - B_2), \end{aligned}$$

where we have put $\mu = \frac{k_0}{k} = \sqrt{1 - \frac{U_0}{E}}.$

From this we obtain

$$A_1 = \tfrac{1}{2}A_2(1 + \mu) + \tfrac{1}{2}B_2(1 - \mu). \tag{197}$$

Analogously, the continuity conditions for $x = l$ give

$$A_2\, e^{2\pi i k_0 l} + B_2\, e^{-2\pi i k_0 l} = A_3\, e^{2\pi i k l},$$
$$A_2\, e^{2\pi i k_0 l} - B_2\, e^{-2\pi i k_0 l} = \frac{1}{\mu} A_3\, e^{2\pi i k l},$$

from which we get

$$A_2 = \frac{\mu + 1}{2\mu} A_3\, e^{2\pi i (k - k_0) l},$$
$$B_2 = \frac{\mu - 1}{2\mu} A_3\, e^{2\pi i (k + k_0) l}.$$

Substituting these expressions into (197), we obtain the following relation between A_1 and A_3:

$$A_1 = \frac{A_3}{4\mu} e^{2\pi i k l}[(\mu + 1)^2 e^{-2\pi i k_0 l} - (\mu - 1)^2 e^{2\pi i k_0 l}]. \qquad (198)$$

From this we obtain τ by means of (196), or more conveniently its reciprocal:

$$\frac{1}{\tau} = \frac{|A_1|^2}{|A_3|^2} = \frac{1}{16|\mu|^2} |(\mu + 1)^2 e^{-2\pi i k_0 l} - (\mu - 1)^2 e^{2\pi i k_0 l}|^2. \qquad (199)$$

In order to calculate this expression, it is better to treat the two cases $E > U_0$ and $E < U_0$ separately:

Case (a): $E > U_0$. In this case k_0 and μ are real, and hence the right-hand member of (199) may be written

$$\begin{aligned}\frac{1}{\tau} &= \frac{1}{16\mu^2} [(\mu + 1)^2 e^{-2\pi i k_0 l} - (\mu - 1)^2 e^{2\pi i k_0 l}] \\ &\qquad \times [(\mu + 1)^2 e^{2\pi i k_0 l} - (\mu - 1)^2 e^{-2\pi i k_0 l}] \\ &= \frac{1}{16\mu^2} [(\mu + 1)^4 + (\mu - 1)^4 - 2(\mu^2 - 1)^2 \cos 4\pi k_0 l]. \qquad (199')\end{aligned}$$

From this equation we see that with μ fixed, the transmission coefficient τ varies in a periodic manner with a variation of l (thickness of the barrier). For $\cos 4\pi k_0 l = 1$ or

$$l = \frac{n}{2k_0} = \frac{n\lambda_0}{2} \qquad (n \text{ integer}), \qquad (200)$$

the square bracket attains its maximum value equal to $16\mu^2$, and hence

$$\tau = \tau_{\max} = 1.$$

For $\cos 4\pi k_0 l = -1$ and hence for

$$l = \frac{(n + \frac{1}{2})}{2k_0} = \left(n + \frac{1}{2}\right)\frac{\lambda_0}{2},$$

we have instead

$$\tau = \tau_{\min} = \frac{4\mu^2}{(1 + \mu^2)^2}.$$

τ always lies between these two limits. As can be seen from the last formula, τ no longer vanishes and hence there is always a certain probability that a particle traverses the barrier. This probability becomes a certainty if the thickness of the barrier contains an integral number of half-wavelengths, as is indicated by (200); note the analogy of this phenomenon with the transmission of light through a thin plate. But in every other case $\tau < 1$, and hence there is a certain probability $1 - \tau$ that the particle will be repelled, in contrast to what would result from classical mechanics.

CASE (b): $E < U_0$. In this case, which is the more interesting one as far as applications are concerned, k_0 and hence μ are imaginary, so that we may write

$$k_0 = ik' = i\sqrt{\frac{2m}{h^2}(U_0 - E)}, \qquad \mu = i\mu' = i\sqrt{\frac{U_0}{E} - 1},$$

with k' and μ' real. Therefore (199) becomes

$$\begin{aligned}\frac{1}{\tau} &= \frac{1}{16\mu'^2}[(1 + i\mu')^2 e^{2\pi k'l} - (1 - i\mu')^2 e^{-2\pi k'l}] \\ &\qquad \times [(1 - i\mu')^2 e^{2\pi k'l} - (1 + i\mu')^2 e^{-2\pi k'l}] \\ &= \frac{1}{16\mu'^2}[-(1 + i\mu')^4 - (1 - i\mu')^4 + 2(1 + \mu'^2)^2 \cosh 4\pi k'l] \quad (201) \\ &= 1 + \frac{(1 + \mu'^2)^2}{8\mu'^2}(\cosh 4\pi k'l - 1) = 1 + \frac{(1 + \mu'^2)^2}{4\mu'^2}\sinh^2 2\pi k'l.\end{aligned}$$

We see that once μ' has been fixed, the right-hand member steadily increases with increasing l, and hence τ steadily decreases with an increase of the barrier thickness (it becomes negligible when l becomes large compared with $1/k'$). However, we see from the formula that τ does not vanish (of course, as long as U_0/E and l remain finite), so that there is always a certain probability of penetrating the barrier and of continuing beyond indefinitely, *even for a particle with energy below* U_0.

Here apparently the same difficulty occurs which was mentioned in connection with the potential step, inasmuch as if we were to imagine the particle as actually moving continuously in the ordinary sense, we should be led to conclude that the particle has traversed a region in which its kinetic energy must be negative. But in reality, as we know, we are not entitled to speak of the motion of a particle between two successive observations. And if an observation of position made us find the particle in the region where $U > E$, this would imply no paradox at all, since (as has already been observed) the very fact that we observed its position altered the value of its energy.

The optical analogue of the phenomenon considered here is the following. It is known that the special luminous disturbance which exists in the second medium during total reflection may be, so to speak, captured and transformed into ordinary radiation, by placing, at a very small distance from the reflecting surface, the surface of another medium equal to the one from which the light emerges. Thus if AMN (Fig. 34) is a glass prism on whose hypotenuse MN the total reflection takes place, upon approaching to within a very small distance (of the order of a wavelength) with a second glass prism $A'M'N'$, we see that part of the incident light passes into that prism and emerges from the face $A'N'$. This phenomenon occurs despite the fact that in the passage of light from the first prism to the air layer there results (according to geometrical optics) an imaginary angle of refraction. The air layer $MNM'N'$ is perfectly analogous to the potential barrier considered above.[6]

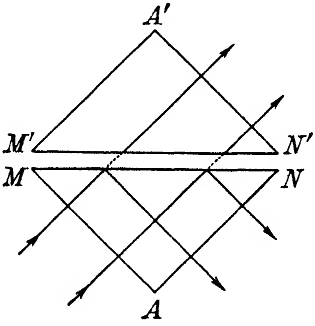

Fig. 34

As far as a direct experimental verification of this result of wave mechanics is concerned, it must be remembered that this seems practically impossible, owing to the fact that the thickness l of the potential barrier would have to be of the order of a de Broglie wavelength, which even for the slowest cathode rays is always very short. Nevertheless, as we shall see in the next few sections, there exist experimental facts which may be interpreted as an indirect confirmation of this remarkable phenomenon.

---

[6] There is, however, the difference that with de Broglie waves the phenomenon occurs even at normal incidence, but not with light.

41. Note on the theory of cold emission of electrons. It has been found experimentally that upon application of a strong electric field X, in vacuum, to the surface of a metal, the field being directed toward the metal, electrons whose number is proportional to the quantity $e^{-b/X}$, where b is a constant, will be induced to leave the metal. The considerations of the preceding section permit a simple explanation of this phenomenon which is partly quantitative.

It is known that in order to remove an electron from a metal, a certain amount of energy U_0 is required for the process of crossing the surface. Therefore we may say that the metallic surface represents a potential step of height U_0, for the electrons so that, if we draw the x-axis perpendicular to the surface of the metal, the potential U of the forces acting upon an electron has the shape of the line $AOBC$ in Fig. 35: it is the wall OB which prevents the electrons (whose energy is less than U_0) from leaving the metal.

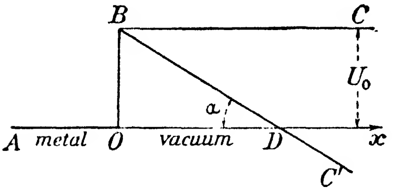

Fig. 35

When the electric field is applied to the outside of the conductor, the part BC of the potential is replaced by the sloped line BC' (whose slope $\tan\alpha$ represents the force eX). Thus, in place of the potential step, a potential "barrier" OBD is formed which, according to classical mechanics, would be just as insurmountable for the electrons. But instead, as we have seen in the previous section, wave mechanics permits some of the electrons to penetrate this barrier and leave the metal.

Fowler and Nordheim[7] have developed a theory of this phenomenon on the foregoing basis, accounting for it in a satisfactory manner, even quantitatively. We confine ourselves here to pointing out that the formulas of the preceding section permit us to foresee that the intensity of the emitted electron current will be expressed, as a function of the field X, by a law of the indicated type. In fact, Fig. 35 shows that the width OD of the barrier is given by

$$OD = \frac{U_0}{\tan\alpha} = \frac{U_0}{eX},$$

---

[7] *Proc. Roy. Soc.* **A 119,** 173 (1928).

and that (ignoring the difference in shape of the barrier, which does not modify the aspect of the qualitative argument) if we identify OD with the thickness of the barrier considered in the last section, the quantity $4\pi k'l$ occurring in formula (201) becomes

$$4\pi k'l = 4\pi\sqrt{\frac{2m}{h^2}(U_0 - E)}\,\frac{U_0}{eX} = \frac{b}{X}, \tag{202}$$

where the portion independent of X is designated by b. Substituting values which occur in practical cases, we find that $4\pi k'l$ comes out at least of the order of 100. Hence, applying formula (201), we may replace cosh $4\pi k'l$ by $e^{4\pi k'l}$ and may neglect unity compared with this term. The formula is thus reduced to

$$\frac{1}{\tau} = \frac{(1 + \mu'^2)^2}{8\mu'^2}\,e^{b/X}, \tag{203}$$

which yields for transmission coefficient τ, and hence for the number of electrons emitted, the proper expression suggested by experimental data.

42. Particle between two potential barriers. Let us now consider the case in which the potential has the shape of Fig. 36, which may be considered to be built up of two symmetric potential barriers of height U_0, enclosing a central "valley"; the bottom of the latter may also be at a lower level than the external potential, as in the figure.

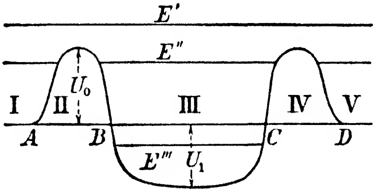

Fig. 36

From the point of view of classical mechanics, three cases must be distinguished. If $E > U_0$ (for example, level E'), a particle will traverse the entire double barrier; if $U_0 > E > 0$ (for example, level E''), three types of clearly distinct motions are possible; (1) the particle comes from the left, will be repelled by the first barrier, and turn back; (2) the particle comes from the right and executes a similar motion along the positive x-axis; or (3) the particle finds itself between the two barriers and then performs periodic oscillations. If $0 > E > -U_1$ (level E'''), only the oscillatory motion between the two barriers is possible.

In order to study the problem according to wave mechanics, it is convenient again to idealize the potential by giving it the shape of Fig. 37. We then apply the same procedure as in §40, solving the Schrödinger equation separately for each of the five regions into which the points A, B, C, and D divide the x-axis, and imposing the condition that u and du/dx be continuous at these points, which establishes a relation between the constants of the various regions. The result obtained may also be extended qualitatively to the case of Fig. 36. Its most remarkable result applies to the case where $U_0 > E > 0$ (levels of the type E'').

For a brief outline of this result, we shall start out with the obvious observation that if the two barriers became infinitely high, a particle in the central region would be subject to the conditions examined in §38 (particle on a line segment), and hence its energy could take on only certain discrete values E_1, E_2, . . . ; in other words, the energy would be "quantized." But if we consider barriers of finite height, these no longer represent insurmountable obstacles for the particle; the energy is no longer quantized, and may assume any value. We find, however, the following result: ψ is evidently of an oscillatory nature in the central region (III) as well as in the two external regions (I, V), whereas it is of the exponential type in regions II and IV. If the energy has any arbitrary value, the amplitude of ψ in the central region will in *general* be considerably smaller than in the outer regions, because the exponential portions which join it approach the x-axis toward the interior. Consequently, the particle has a very small probability of finding itself in the central region, and if found there it has a strong tendency to leave. Only in the case where the energy lies *close* to one of the

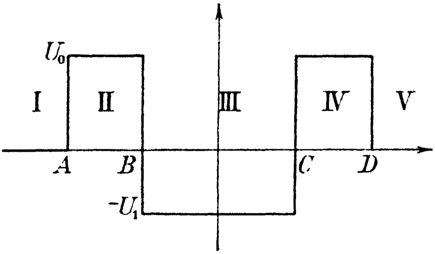

Fig. 37

values E_1, E_2, . . . (which in the case of infinitely high barriers are the only permitted values) does the amplitude in the central region become much larger than in the outer regions; that is, the particle has a large probability of being found inside. Hence we can say that instead of rigorous quantization, we are dealing with *quasi quantization*, meaning that the discrete levels E_1, E_2, . . . are no longer represented by sharp horizontal lines but by narrow shaded bands. They also no longer represent the only permitted values of E; instead, they represent the values of E which give to the particle a considerable probability of being found inside the "valley" limited by the two barriers.

For energy levels below 0 (type E''') we have ordinary quantization and ψ is practically confined to the central region. On the other hand, the levels higher than the barriers, that is, of type E', are not quantized at all and correspond to a ψ which is sinusoidal everywhere.

43. Note on the theory of alpha-particle emission. The considerations of the last section constitute the basis for a remarkable theory devised by Gamow, and independently by Gurney and Condon, in order to account for the spontaneous emission of α-particles from the nuclei of radioactive substances.

This phenomenon gave rise to a serious theoretical difficulty. Indeed, it was possible to determine, by means of experiments of the type which had led Rutherford to the discovery of the atomic nucleus (scattering of α-particles), what force acts upon an α-particle in the neighborhood of a nucleus. It was found that this force is of the Coulomb type and repulsive, as long as the distance from the nucleus is not below a certain limit (of the order of 10^{-11} or 10^{-12} cm). In the cases in which it was experimentally possible to go below that limit (that is, for light elements) it was found that when the nucleus was approached, the repulsion increased less rapidly than was required by Coulomb's law. This effect gave indication that this repulsion was to decrease further and finally to turn into an attraction, as is required for the interior of the nucleus in order to ensure its stability. The potential $U(r)$ will therefore be given by a curve of the shape shown in Fig. 38, with a maximum U_0 corresponding to the distance r_0 at which the force becomes attractive. The α-particles which are part of the nucleus are within the region corresponding to the potential "valley." The case of greatest interest is the one of uranium I(U I). For this element it has been possible to make an experimental determination of the curve only down to distances of about 3×10^{-12} cm (solid portion in the figure), but this is sufficient to exhibit the paradox mentioned previously, which is as follows: the α-particles emitted by uranium I have an energy of about 4 million volts, which is much less (about one half) than the energy corresponding to the highest point M of the curve that was ascertained experimentally. The energy corresponds to α-particles which would have originated, with zero initial velocity, from a distance 6×10^{-12} cm (point N of the curve). Therefore (even independently of any hypothesis concerning the shape of the curve in the portion not established by experiment) it remains unexplained, from the classical point of view, how particles having an energy less than the maximum, U_0, of the potential barrier surrounding the nucleus have succeeded in issuing from the nucleus.

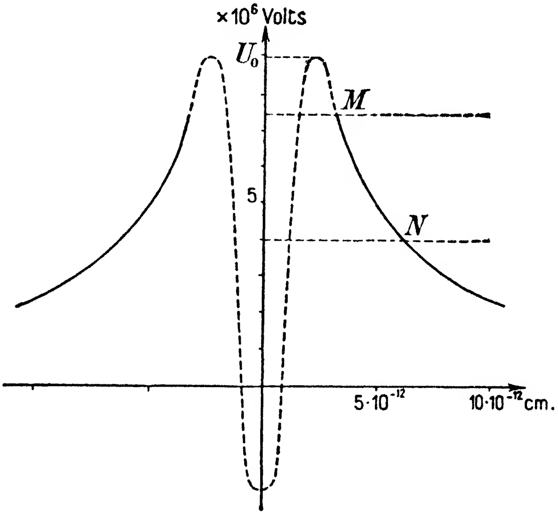

Fig. 38

From the point of view of wave mechanics, however, this difficulty disappears, because, as we have seen, a particle may occasionally penetrate a potential barrier higher than its own energy. The α-particles of the nucleus behave qualitatively like the particles of the one-dimensional problem studied in the last section (from which the present problem differs by having spherical symmetry, which does not affect anything essential); they will *ordinarily* remain in the interior of the nucleus, with certain quasi-quantized energy levels, but will have a certain probability of leaving it, passing, as it were, through the barrier but keeping the same energy. This effect is in perfect accord with the experimental fact that the α-particles emitted from a given substance all have certain well-defined energies (that is, they have a "line spectrum") which just represent the energy levels in question. In practice these energy levels turn out to be almost perfectly quantized, the width of the bands being altogether negligible. A quantitative determination of these energy levels would require an exact knowledge of the potential curve, of which, however, we know only the qualitative behavior for the interior. Fortunately, this exact knowledge is not necessary for a calculation of the transmission coefficient of the barrier, that is, the probability that a particle will penetrate it, upon which probability evidently depends the mean life of the radioactive element considered.

In this way a simple relation[8] between the mean life of the element and the velocity of the α-particles emitted was found. This relation has already been observed empirically by Geiger and Nuttall for all radioactive substances emitting α-particles.

---

[8] See No. 29 of the Bibliography.

CHAPTER 8

Three-dimensional Problems

44. Particle subject to no forces. In this chapter we shall study the wave mechanics of a particle, removing the restriction of the preceding chapter that all quantities depend only on x and t; hence three spatial coordinates will now enter, in addition to the time.

The first problem treated here is that of the free particle in space with no forces present. Its most general ψ may be obtained (see §29) as a sum of *monochromatic components* each of which corresponds to a definite value of the energy E. Hence we shall look for these monochromatic solutions, of frequency $\nu = E/h$—that is, of the form (128) of §27.

Again let us take up equation (131′), which $u(x, y, z)$ satisfies, writing it with Δ expressed explicitly and with $U = 0$:

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} + \frac{8\pi^2 m}{h^2} E u = 0.$$

This well-known equation may be integrated by the method of separation of variables, that is, by letting

$$.u(x, y, z) = X(x) Y(y) Z(z). \tag{204}$$

The equation then becomes (after we divide through by XYZ):

$$\frac{X''}{X} + \frac{Y''}{Y} + \frac{Z''}{Z} + \frac{8\pi^2 m}{h^2} E = 0.$$

If we observe that each of the first three terms depends on only one of the coordinates, we realize that, in order to have the equation satisfied for any x, y, z, we must have

$$\frac{X''}{X} = c_1, \qquad \frac{Y''}{Y} = c_2, \qquad \frac{Z''}{Z} = c_3, \tag{205}$$

c_1, c_2 and c_3 being constants connected through the relation

$$c_1 + c_2 + c_3 + \frac{8\pi^2 m}{h^2} E = 0. \tag{206}$$

The first equation of (205) has a general solution of the type

$$X = a_1 \, e^{\sqrt{c_1} x} + b_1 \, e^{-\sqrt{c_1} x},$$

which would become infinite either for $x = +\infty$ or for $x = -\infty$, if c_1 were positive. However, since it is required that u be finite everywhere, we conclude that $c_1 \leq 0$, so that in the expression for X the exponents will become imaginary (or zero) and X will be of an oscillatory character (or constant). The same arguments hold for Y and for Z. We may therefore put (indicating three new constants by k_x, k_y, k_z)

$$c_1 = -4\pi^2 k_x^2, \qquad c_2 = -4\pi^2 k_y^2, \qquad c_3 = -4\pi^2 k_z^2.$$

Thereupon (206) becomes, after we set $\sqrt{k_x^2 + k_y^2 + k_z^2} = k$,

$$k^2 = \frac{2mE}{h^2}, \tag{207}$$

and the general solutions of equations (205) may be written

$$\left.\begin{aligned} X &= a_1 \, e^{2\pi i k_x x} + b_1 \, e^{-2\pi i k_x x}, \\ Y &= a_2 \, e^{2\pi i k_y y} + b_2 \, e^{-2\pi i k_y y}, \\ Z &= a_3 \, e^{2\pi i k_z z} + b_3 \, e^{-2\pi i k_z z}. \end{aligned}\right\}$$

Each of these functions is of the same form as the u found for the one-dimensional problem [§35, formula (149)]. As we saw for that case, we are able to consider only the first terms (the second terms are obtained by changing the sign of k_x, k_y, k_z), and then u takes on the form

$$u = A \, e^{2\pi i (k_x x + k_y y + k_z z)}, \tag{208}$$

and ψ becomes

$$\psi = A \, e^{2\pi i (k_x x + k_y y + k_z z - \nu t)}, \tag{209}$$

where [see (207)]

$$\nu = \frac{1}{h} E = \frac{h}{2m} k^2. \tag{207'}$$

It is convenient (as in §15) to introduce the vector **k** with components k_x, k_y, k_z (and hence of modulus k) and the vector **r** originating at the origin of the coordinate axes and ending at the point P where ψ is being calculated, **r** therefore having components x, y, z. Now the preceding formula is written in invariant form, that is, independent of the axes:

$$\psi = A\, e^{2\pi i(\mathbf{k}\cdot\mathbf{r} - \nu t)}. \tag{210}$$

This formula is identical with (78′) of §15. As we have seen, (78′) represents a plane wave train having **k** for its "propagation vector," and hence a wavelength $\lambda = 1/k$ and a direction of propagation given by the vector **k**. Therefore the phenomenon is physically the same as the one studied in §35 but is now referred to arbitrary axes. Consequently, this wave train represents a particle whose momentum **p** is exactly determined ($\mathbf{p} = h\mathbf{k}$), and whose energy is exactly defined as

$$E = h\nu = \frac{h^2}{2m} k^2,$$

while its position is totally indeterminate. If **k** and ν in (210) are expressed in terms of **p**, the expression becomes

$$\psi = A\, e^{\frac{2\pi i}{h}\left(\mathbf{p}\cdot\mathbf{r} - \frac{p^2}{2m}t\right)}. \tag{210′}$$

We now proceed to consider the most general solutions, obtained by superimposing infinite wave trains like the one above, but with different propagation vectors; these are the solutions analogous to (79) of §15. Expressing **k** in terms of **p** in (79) and calling $\varphi_0(\mathbf{p})$ the function $h^{3/2}A(\mathbf{k})$ (proportional to the amplitude of the wave train of momentum **p**), we obtain the following expression for ψ:

$$\psi = h^{-3/2}\iiint \varphi_0(\mathbf{p})\, e^{\frac{2\pi i}{h}\left(\mathbf{p}\cdot\mathbf{r} - \frac{p^2}{2m}t\right)} dp_x\, dp_y\, dp_z, \tag{211}$$

or else, putting

$$\varphi(\mathbf{p},t) = \varphi_0(\mathbf{p})\, e^{-\frac{2\pi i}{h}\frac{p^2}{2m}t}, \tag{212}$$

$$\psi = h^{-3/2}\iiint \varphi(\mathbf{p},t)\, e^{\frac{2\pi i}{h}\mathbf{p}\cdot\mathbf{r}} dp_x\, dp_y\, dp_z. \tag{213}$$

For $t = 0$, ψ becomes

$$\psi_0 = h^{-3/2}\iiint \varphi_0(\mathbf{p})\, e^{\frac{2\pi i}{h}\mathbf{p}\cdot\mathbf{r}} dp_x\, dp_y\, dp_z. \tag{213′}$$

Remembering that A of (79) is given by (80), we see that φ_0 is obtained from the initial wave function ψ_0 by the formula

$$\varphi_0(\mathbf{p}) = h^{-3/2}\int\int\int\psi_0(\mathbf{r})\, e^{-\frac{2\pi i}{h}\mathbf{p}\cdot\mathbf{r}} dx\, dy\, dz; \tag{214}$$

hence we get φ from ψ by means of

$$\varphi(\mathbf{p},t) = h^{-3/2}\int\int\int\psi(\mathbf{r},t)\, e^{-\frac{2\pi i}{h}\mathbf{p}\cdot\mathbf{r}} dx\, dy\, dz. \tag{215}$$

Note the analogy between expressions (213) and (215), which may be considered each other's inverse and in which the functions φ and ψ play symmetrical roles.

Recalling the principle of superposition, we may interpret the solution (213) in the following manner. When the state of the particle is represented by (213), if we perform a measurement of the momentum, there is a probability

$$|\varphi_0|^2 dp_x\, dp_y\, dp_z = |\varphi|^2 dp_x\, dp_y\, dp_z$$

of finding the components of the momentum lying between p_x and $p_x + dp_x$, p_y and $p_y + dp_y$, p_z and $p_z + dp_z$. The function $\varphi(\mathbf{p})$, as far as the measurements of momentum are concerned, has the same significance which ψ has with respect to measurements of position.

It is hardly necessary to point out that if ψ_0 given by (213′) is different from zero only in a limited region of space, it will represent a wave packet which moves approximately with uniform rectilinear motion; that is, it moves like a particle in classical mechanics, spreading out gradually in three dimensions, however, as may be seen by generalizing the calculation of §36.

In this case too, it may be verified that wave mechanics contains within itself the uncertainty principle (thus generalizing what has been said in §36 in connection with the one-dimensional problem); it suffices to take expressions (83) of Chapter 5 and to express **k** in terms of **p**, thus obtaining

$$\left.\begin{aligned} \Delta x\, \Delta p_x &\geq \frac{h}{4\pi}, \\ \Delta y\, \Delta p_y &\geq \frac{h}{4\pi}, \\ \Delta z\, \Delta p_z &\geq \frac{h}{4\pi}, \end{aligned}\right\} \tag{216}$$

which embody the uncertainty principle for a particle in space.

45. Particle in a rectangular box. In analogy to what we did in the one-dimensional case (§38), we may now briefly consider the case of a particle constrained to remain within a rectangular box of sides a, b, c, with perfectly reflecting walls. No forces are present in the interior of the box. Let us first look for a stationary solution corresponding to a given value of E or to a single frequency, being guided by the analogy with the problem of light or sound waves in a box with reflecting walls.

It is clear that a system of plane waves propagating in a direction with direction cosines α, β, γ, upon reflection will give rise to a system of waves characterized by the direction cosines $\pm\alpha$, $\pm\beta$, $\pm\gamma$, from the walls (where the signs may be combined in all eight possible ways). Therefore we cannot use a solution like (209) by itself but instead must take a sum of eight terms, of the type

$$b\, e^{-\frac{2\pi i}{h}Et}\, e^{2\pi i(\pm k_x x \pm k_y y \pm k_z z)}. \tag{217}$$

But now we must have $u = 0$ along the walls for all t. This condition leads to a restriction of the arbitrariness of the propagation vector **k**. In fact, let us suppose that two walls correspond to the planes $x = 0$, $x = a$, so that the above-mentioned sum may be written in the form

$$u = f_1(y, z, t)\, e^{2\pi i k_x x} + f_2(y, z, t)\, e^{-2\pi i k_x x};$$

we see immediately that to satisfy the condition $u = 0$ for $x = 0$ and for $x = a$ for any y, z, t, we must have $2\pi k_x a = n_1\pi$, with n_1 an integer, and similarly for k_y and k_z. Hence

$$k = \frac{1}{2}\sqrt{\frac{n_1^2}{a^2} + \frac{n_2^2}{b^2} + \frac{n_3^2}{c^2}}; \tag{218}$$

and since $k = 1/\lambda$, it may be concluded that the only wavelengths which may give rise to stationary waves are the ones expressible by the formula

$$\lambda_{n_1 n_2 n_3} = \frac{2}{\sqrt{\frac{n_1^2}{a^2} + \frac{n_2^2}{b^2} + \frac{n_3^2}{c^2}}}, \tag{219}$$

and hence the only possible values for E are those given by

$$E_{n_1n_2n_3} = \frac{h^2}{8m}\left(\frac{n_1^2}{a^2} + \frac{n_2^2}{b^2} + \frac{n_3^2}{c^2}\right). \tag{220}$$

Thus the system has a triple infinity of discrete energy levels characterized by three quantum numbers n_1, n_2, n_3.

Analogously in acoustics, a rectangular cavity may resonate only for certain definite frequencies (*proper frequencies*), to which there correspond wavelengths given exactly by (219).

In the more general case of a particle enclosed in a cavity of any shape, the problem is always analogous to the acoustical problem of finding the *proper frequencies*, or resonance frequencies, for which stationary waves may be set up in the interior of the cavity; this problem always leads to discrete energy levels.

46. Central forces (general). Let us now consider a particle subject to the action of a central force. It will evidently be convenient to use the polar coordinates r, θ, φ, having their origin at the center of force, and the potential U will be a function of r alone.

The Schrödinger equation (127), with the explicit expression for the operator Δ in terms of polar coordinates, is written

$$\frac{1}{r}\frac{\partial^2(ru)}{\partial r^2} + \frac{1}{r^2 \sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial u}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2 u}{\partial\varphi^2} + \frac{8\pi^2 m}{h^2}(E - U)u = 0. \tag{221}$$

Let us try to separate the variable r from θ and φ by setting

$$u = R(r)Y(\theta, \varphi). \tag{222}$$

Substituting into the preceding equation and multiplying through by r^2/RY, we obtain

$$\frac{r}{R}\frac{d^2(rR)}{dr^2} + \frac{1}{Y\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial Y}{\partial\theta}\right) + \frac{1}{Y\sin^2\theta}\frac{\partial^2 Y}{\partial\varphi^2} + \frac{8\pi^2 m}{h^2}r^2(E - U) = 0.$$

In this equation the first and last terms depend only on r, the other two terms only on θ and φ. Hence the equation is separated into two:

$$\frac{1}{Y \sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial Y}{\partial\theta}\right)+\frac{1}{Y\sin^2\theta}\frac{\partial^2 Y}{\partial\varphi^2}=-C, \tag{223}$$

$$\frac{r}{R}\frac{d^2(rR)}{dr^2}+\frac{8\pi^2 m}{h^2}r^2(E-U)=C, \tag{224}$$

where C is a constant. The first equation does not contain the function $U(r)$ and is therefore common to all problems with central forces. Hence in (222) the factor $Y(\theta, \varphi)$ is independent of the law according to which the force varies as a function of r, whereas the factor $R(r)$ does depend on that law and is therefore different in various problems. In this section we shall study equation (223), which holds for all problems with central forces. It may be written

$$\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial Y}{\partial\theta}\right)+\frac{1}{\sin^2\theta}\frac{\partial^2 Y}{\partial\varphi^2}+CY=0, \tag{223'}$$

and it may be shown that the equation has solutions which are finite, continuous, and single-valued for all directions only if

$$C=l(l+1), \tag{225}$$

with $l = 0, 1, 2, \ldots$. The integer l is called *azimuthal quantum number* because it corresponds to the azimuthal quantum number of the Bohr-Sommerfeld theory. With the expression (225) for C, we recognize in (223′) the differential equation for spherical harmonics of order l.

Upon further separation of the variables θ and φ by the substitution

$$Y(\theta,\varphi)=\Theta(\theta)\cdot\Phi(\varphi), \tag{226}$$

the equation again splits into two:

$$\frac{d^2\Phi}{d\varphi^2}+\lambda\Phi=0, \tag{227}$$

$$\frac{1}{\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right)+\left(C-\frac{\lambda}{\sin^2\theta}\right)\Theta=0, \tag{228}$$

where λ is a constant.

Equation (227) gives

$$\Phi = \text{const. } e^{\pm i\sqrt{\lambda}\varphi}. \tag{229}$$

Since Φ must be periodic in φ with period 2π (otherwise u would not be single-valued for every point of space), $\sqrt{\lambda}$ must be equal to an integer[1] m; that is,

$$\lambda = m^2. \tag{230}$$

Then, if we determine the constant so as to make

$$\int_0^{2\pi} |\Phi|^2 \, d\varphi = 1, \tag{231}$$

(229) becomes

$$\Phi = \frac{1}{\sqrt{2\pi}} e^{im\varphi}, \tag{229'}$$

where the exponent has been written without the double sign, with the understanding that m may be positive, zero, or negative. The integer m (with its sign) is called *magnetic quantum number.*

In order to study (228) it is convenient to take for our variable $\cos\theta$, which we shall indicate by x. Then the equation becomes, with (230) taken into account,

$$(1 - x^2)\frac{d^2\Theta}{dx^2} - 2x\frac{d\Theta}{dx} + \left(C - \frac{m^2}{1 - x^2}\right)\Theta = 0. \tag{232}$$

The eigenvalues and eigenfunctions of this equation are investigated by a method analogous to that followed in §39 for the oscillator. The singular points of the equation are $x = \pm 1$ and $x = \infty$, as may be readily recognized by dividing through by $(1 - x^2)$ (see §16); and we note that we are dealing with singularities of the Fuchsian type. Since in our case, x ranges from -1 to $+1$, we are interested only in the singularities at the endpoints of this interval. By the method of the indicial equation, we find that in the neighborhood of $x = +1$, Θ may be put in the form

$$(1 - x)^{\pm\frac{m}{2}} f_1(1 - x),$$

---

[1] We shall adopt m in order to conform to universal usage, although m also designates the mass of the particle.

where f_1 is a holomorphic function which does not vanish for $1 - x = 0$. Of the two forms, the one with the negative exponent has a pole for $1 - x = 0$ and hence is to be discarded. Therefore $\Theta = (1 - x)^{\frac{|m|}{2}} f_1(1 - x)$. Similarly, the other singular point gives $\Theta = (1 + x)^{\frac{|m|}{2}} f_2(1 + x)$, with f_2 holomorphic and nonzero at $1 + x = 0$. Since these two expressions must represent two different expansions of the same branch of the continuous function Θ, we must be able to write

$$\Theta = (1 - x^2)^{\frac{|m|}{2}} P(x), \tag{233}$$

with

$$P(x) = \sum_{r=0}^{\infty} a_r x^r. \tag{234}$$

Substituting (233) into (232), we find for P the equation

$$(1 - x^2)P'' - 2(|m| + 1)xP' + [C - |m|(|m| + 1)]P = 0; \tag{235}$$

and when trying to satisfy (235) with the series (234), we find for a_r the recursion relation

$$\begin{aligned}(r + 1)(r + 2)a_{r+2}& \\ &= [r(r - 1) + 2(|m| + 1)r - C + |m|(|m| + 1)]a_r.\end{aligned} \tag{236}$$

Hence we may take for P a series in even or in odd powers, with arbitrary first coefficient.

The condition that P be finite over the whole interval under consideration is certainly met if one of the coefficients, $a_{\gamma+2}$ for example, vanishes together with all succeeding ones, and the series P reduces to a polynomial of degree γ. The condition for $a_{\gamma+2} = 0$ (a_γ being $\neq 0$) is, as may be seen from (235), that

$$\begin{aligned}C = \gamma(\gamma - 1) + 2(|m| + 1)\gamma + |m|(|m| + 1)& \\ &= (\gamma + |m|)(\gamma + |m| + 1);\end{aligned}$$

and if we put

$$\gamma + |m| = l \tag{237}$$

(from which l comes out to be a nonnegative integer $\geq m$), we obtain (225).

It may also be shown that this condition is not only sufficient but also necessary;[2] that is, if P does not reduce to a polynomial, it cannot satisfy the desired conditions.

Legendre polynomials. Let us first of all consider equation (235) for $m = 0$, in which case Y does not depend on φ and coincides (within a constant factor) with Θ, and the latter with P. With (225) taken into account, the equation is written

$$(1 - x^2)P'' - 2xP' + l(l + 1)P = 0, \tag{238}$$

and the desired solution is a polynomial $P(x)$ of degree l, satisfying the recursion formula (236). These polynomials, which were known for a long time in the theory of spherical harmonics, are defined as follows. Choose $a_1 = 0$ if l is even, and $a_0 = 0$ if l is odd, so that in the first case the polynomial has only even powers, in the second case only odd powers. The other coefficient is chosen in such a way as to make $P_l(1) = 1$. The polynomial $P_l(x)$, defined in this manner, is called *Legendre polynomial* of degree l. When expressed in the form $P_l(\cos\theta)$ it is also called *zonal harmonic* (in fact, because of what has been said above, this is a particular case of spherical harmonics having axial symmetry). It may be shown that all the l roots of the polynomial are real and lie between -1 and $+1$ (endpoints included). Following are the expressions for the first six Legendre polynomials:

$$\begin{aligned} P_0(x) &= 1 & P_1(x) &= x \\ P_2(x) &= \tfrac{3}{2}x^2 - \tfrac{1}{2} & P_3(x) &= \tfrac{5}{2}x^3 - \tfrac{3}{2}x \\ P_4(x) &= \tfrac{35}{8}x^4 - \tfrac{15}{4}x^2 + \tfrac{3}{8} & P_5(x) &= \tfrac{63}{8}x^5 - \tfrac{35}{4}x^3 + \tfrac{15}{8}x. \end{aligned}$$

Since these expressions are eigenfunctions of (238), they constitute a system of orthogonal functions in the interval $(-1, +1)$. They are not normalized, however, since

$$\int_{-1}^{+1} P_l^2(x)\,dx = \frac{2}{2l + 1}. \tag{239}$$

These expressions may also be defined by means of the lth derivative of the expression $(x^2 - 1)^l$. In fact, we have

$$P_l(x) = \frac{1}{2^l l!}\frac{d^l}{dx^l}(x^2 - 1)^l. \tag{240}$$

---

[2] See Bechert, *Ann. d. Physik* **83**, 906 (1927).

Another one of their important properties is expressed by the recursion formula which relates three successive polynomials:

$$P_l = \frac{2l-1}{l} xP_{l-1} - \frac{l-1}{l} P_{l-2}, \qquad (241)$$

which permits us to calculate successively all the Legendre polynomials starting with P_0 and P_1.

Associated Legendre functions. Let us proceed now to a consideration of (235) without the restriction $m = 0$. It may be written, taking (225) into account, as

$$(1 - x^2)P'' - 2(|m| + 1)xP' + [l(l + 1) - |m|(|m| + 1)]P = 0, \qquad (242)$$

with $|m| \leq l$.

Its eigenfunctions may be obtained by means of the recursion formula (236), but it may also be noted that if we take the derivative of the last equation with respect to x, we obtain an equation of the same form, in which instead of P, P', P'' there occur P', P'', P''', respectively, and where instead of $|m|$ there is $(|m| + 1)$, so that if P_l satisfies (242) for $|m| = 0$, as we have seen, P_l' will satisfy it for $|m| = 1$, P_l'' for $|m| = 2$, and so on. Therefore (242) is satisfied by $\frac{d^{|m|}P_l}{dx^{|m|}}$, which is a polynomial of degree $l - |m| = \gamma$, and hence (232) is satisfied by taking Θ equal (or proportional) to

$$P_l^m(x) = (1 - x^2)^{\frac{|m|}{2}} \frac{d^{|m|}P_l}{dx^{|m|}}.$$

These functions are called *associated Legendre functions.* They are of course orthogonal over the interval $(-1, +1)$ but are not normalized, since

$$\int_{-1}^{+1} (P_l^m)^2 dx = \frac{(l + |m|)!}{(l - |m|)!} \frac{2}{2l + 1}.$$

Thus, in order that Θ also be normalized (with respect to the variable x), it will suffice to take for its expression

$$\Theta_{lm} = \sqrt{\frac{(l - |m|)!}{(l + |m|)!} \frac{2l + 1}{2}} P_l^m(x)$$
$$= \sqrt{\frac{(l - |m|)!}{(l + |m|)!} \frac{2l + 1}{2}} \sin^{|m|} \theta \frac{d^{|m|}P_l(\cos \theta)}{d(\cos \theta)^{|m|}}, \qquad (243)$$

and we then get $$\int_0^\pi \Theta_{lm}^2 \sin\theta\, d\theta = 1. \tag{244}$$

To summarize, to each eigenvalue $C = l(l+1)$ of (223′) there correspond $2l+1$ eigenfunctions $Y_{lm}(\theta, \varphi)$ (with $m = -l, -l+1, \ldots 0, 1, \ldots l$), given by (226), (229′), (243), that is, by

$$Y_{lm} = \frac{1}{N_{lm}} e^{im\varphi} \sin^{|m|}\theta \frac{d^{|m|}P_l(\cos\theta)}{d(\cos\theta)^{|m|}}, \tag{245}$$

where we have placed in evidence the normalization factor

$$\frac{1}{N_{lm}} = \frac{1}{\sqrt{2\pi}} \sqrt{\frac{(l-|m|)!}{(l+|m|)!} \frac{2l+1}{2}}. \tag{246}$$

Y_{lm} thus normalized has the property, as a consequence of (231) and (244), that

$$\int\int |Y_{lm}|^2 \sin\theta\, d\theta\, d\varphi = 1, \tag{246′}$$

where the integral is extended over the whole spherical surface.

These functions are particular *spherical (surface) harmonics* of order l. Of these, the harmonic corresponding to $m = 0$ reduces to

$$Y_{l0} = \frac{1}{N_{l0}} P_l(\cos\theta); \tag{247}$$

it does not depend on φ, and hence has axial symmetry.

Furthermore, we note that to the eigenvalue zero ($l = 0$) there corresponds only the eigenfunction

$$Y_{00} = \frac{1}{N_{00}} P_0 = \frac{1}{N_{00}} = \frac{1}{2\sqrt{\pi}}, \tag{248}$$

that is, a constant. Hence u in this case depends only on r (spherical symmetry).

We shall list here, for the convenience of the reader, the explicit expressions of the spherical harmonics corresponding to the first four values of l which occur most often in atomic mechanics. We shall let m take on only the values 0, 1, 2, 3; to obtain the functions corresponding to the negative values of m, it is necessary only to change the sign of the exponent of e.

$$Y_{00} = \frac{1}{2\sqrt{\pi}},$$

$$Y_{10} = \sqrt{\frac{3}{4\pi}} \cos\theta, \qquad Y_{11} = \sqrt{\frac{3}{8\pi}} \sin\theta\, e^{i\varphi},$$

$$Y_{20} = \sqrt{\frac{5}{4\pi}} \left(\frac{3}{2}\cos^2\theta - \frac{1}{2}\right), \qquad Y_{21} = \sqrt{\frac{15}{8\pi}} \sin\theta \cos\theta\, e^{i\varphi},$$

$$Y_{22} = \frac{1}{4}\sqrt{\frac{15}{2\pi}} \sin^2\theta\, e^{2i\varphi},$$

$$Y_{30} = \sqrt{\frac{7}{4\pi}} \left(\frac{5}{2}\cos^3\theta - \frac{3}{2}\cos\theta\right),$$

$$Y_{31} = \frac{1}{4}\sqrt{\frac{21}{4\pi}} \sin\theta(5\cos^2\theta - 1)\, e^{i\varphi},$$

$$Y_{32} = \frac{1}{4}\sqrt{\frac{105}{2\pi}} \sin^2\theta \cos\theta\, e^{2i\varphi}, \qquad Y_{33} = \frac{1}{4}\sqrt{\frac{35}{4\pi}} \sin^3\theta\, e^{3i\varphi}.$$

We should add now, anticipating a result which will be proved in Part III, that the azimuthal quantum number l and the magnetic quantum number m have the following physical significance: the modulus p of the angular momentum of the electron (moment of momentum) with respect to the nucleus is

$$p = \sqrt{l(l+1)}\,\frac{h}{2\pi},$$

and the angular momentum component of the electron along the z-axis is $m(h/2\pi)$.

47. The radial equation in the problem of central forces. We shall now deal with the factor $R(r)$ of (222), which depends on the force law. This factor satisfies equation (224), where C is to be replaced by $l(l+1)$. By setting

$$y(r) = rR(r), \tag{249}$$

we may write this equation

$$\frac{d^2y}{dr^2} + \left[\frac{8\pi^2 m}{h^2}(E - U) - \frac{l(l+1)}{r^2}\right] y = 0. \tag{250}$$

Upon y, we must furthermore impose the conditions that for $r = 0$, R is to remain finite or have a pole of order < 1, so that

$$y(0) = 0,$$

and that for $r \to \infty$, R is to tend to zero not less rapidly than $1/r$, and hence that y is to tend to a finite limit or to zero.

These conditions lead, when the potential $U(r)$ is specified, to the determination of a sequence of eigenvalues for E (in general partly discrete, partly continuous) which represent the energy levels of the system. If they are discrete, they may be designated by an index n (*principal quantum number*) in addition to the index l which already occurs in (250), so that we shall write E_{nl}, and for the corresponding eigenfunctions, R_{nl}. As far as the normalization of these is concerned, we note that $u(r, \theta, \varphi)$ must be normalized in such a manner that

$$\int_0^\infty dr \int_0^\pi d\theta \int_0^{2\pi} |u|^2 r^2 \sin\theta \, d\varphi = 1; \tag{251}$$

since $u = R(r)\Theta(\theta)\Phi(\varphi)$ and since the two last factors are already normalized in accordance with (231) and (244), R must be normalized so that

$$\int_0^\infty R^2(r) r^2 \, dr = 1. \tag{252}$$

We note that although u depends on three quantum numbers (n, l, m), the energy levels depend on only two of them, since the magnetic quantum number m does not enter. Each of these levels is thus a multiplet of order $2l + 1$ (*degeneracy;* see §6), since there are as many independent solutions for u, corresponding to the values which m can take (from $-l$ to $+l$).

Finally, it should be pointed out that in (250) we may also incorporate the term $l(l + 1)/r^2$ into the potential, considering as new potential the function

$$U_1(r) = U(r) + \frac{h^2}{8\pi^2 m} \frac{l(l+1)}{r^2}, \tag{253}$$

whereupon the equation takes the form

$$\frac{d^2y}{dr^2} + \frac{8\pi^2 m}{h^2}(E - U_1)y = 0, \tag{254}$$

identical with the equation for one-dimensional motion under the action of a potential U_1. This form may be formally interpreted by saying that to the force deriving from the potential U there is added the force deriving from the second term of (253), which has its analogue in the centrifugal force of classical mechanics and likewise depends on the angular momentum, that is, on the azimuthal quantum number l.

48. Theory of hydrogenlike systems. The eigenvalues. Let us now apply the theory of central forces to the case of the hydrogen atom and of other so-called "hydrogenlike" systems, formed (see §14, of Part I) of a nucleus of charge Ze and of a single electron (the case of hydrogen corresponds to $Z = 1$). In view of the preponderant mass of the nucleus, we may consider the latter as fixed[3] and therefore the electron to be subject to a central attractive force equal to $Z\,e^2/r^2$. Hence the potential is

$$U = -\frac{Z\,e^2}{r}.$$

Whatever has been said in general in §46 concerning the dependence of ψ on the coordinates θ and φ, may be carried over bodily to the present case, since it is independent of the force law. We must only specialize the radial equation (250), which becomes, with the present expression for U,

$$\frac{d^2y}{dr^2} + \left[\frac{8\pi^2m}{h^2}E + \frac{8\pi^2m}{h^2}\frac{Z\,e^2}{r} - \frac{l(l+1)}{r^2}\right]y = 0. \qquad (255)$$

In order to reduce this equation to a simple form, it is convenient to let the constant term in the brackets equal $\pm 1/r_0^2$, where the plus sign holds if $E > 0$ and the minus sign when $E < 0$, and where r_0 has the dimensions of a length and is given by

$$r_0 = \sqrt{\frac{h^2}{8\pi^2m|E|}}; \qquad (256)$$

furthermore, it is convenient to take for independent variable

$$2\frac{r}{r_0} = x, \qquad (257)$$

---

[3] A more rigorous treatment, taking the motion of the nucleus into account, will be given in §21 of Part III.

whereupon the equation becomes

$$\frac{d^2y}{dx^2} + \left[\pm \frac{1}{4} + \frac{A}{x} - \frac{l(l+1)}{x^2} \right] y = 0, \tag{258}$$

where we have put

$$A = \frac{Z\,e^2}{2r_0|E|} = Z\,e^2 \sqrt{\frac{2\pi^2 m}{h^2|E|}}. \tag{259}$$

The equation has a singularity of the Fuchsian type for $x = 0$ and a non-Fuchsian singularity for $x = \infty$. We are to find the values of the parameter A for which this equation has solutions satisfying the boundary conditions [$y(0) = 0$ and y not infinite for $x \rightarrow \infty$], and the corresponding eigenfunctions. At this point it is perhaps best to treat the two cases $E < 0$ and $E > 0$ separately. In the first case, we must take the minus sign in (258), in the second case the plus sign.[4]

CASE 1: $E < 0$. The asymptotic behavior of the solutions of (258), for $x \rightarrow \infty$, may be obtained by considering that the equation approaches the form

$$\frac{d^2y}{dx^2} - \frac{1}{4} y = 0,$$

of which two fundamental solutions are

$$y = e^{\pm x/2}.$$

Discarding the solution with the plus sign, which tends toward infinity, we are led to search for solutions which at infinity behave as $e^{-x/2}$. This suggests that we put

$$y = e^{-x/2} v(x). \tag{260}$$

With this substitution, the given equation is transformed into the following equation for $v(x)$:

$$v'' - v' + \left[\frac{A}{x} - \frac{l(l+1)}{x^2} \right] v = 0, \tag{261}$$

which for the finite region of x does not have singularities other than the one at $x = 0$ (Fuchsian type). For this singularity the indicial equation is

$$\alpha(\alpha - 1) - l(l+1) = 0,$$

---

[4] For a more complete treatment, see, for example, No. 1 of the Bibliography.

from which $\alpha = -l$, or $\alpha = l + 1$. Discarding the negative solution, which would render v infinite at the origin, we are left with

$$v = x^{l+1}\omega(x), \tag{262}$$

where $\omega(x)$ is of the form

$$\omega = \sum_{s=0}^{\infty} a_s x^s. \tag{263}$$

Substituting (262) into (261), we find that ω must satisfy the equation

$$\omega'' + \left[\frac{2(l+1)}{x} - 1\right]\omega' + \frac{A - (l+1)}{x}\,\omega = 0. \tag{264}$$

Then, substituting (263) and equating the coefficients of each power to zero, we find for a_s the recursion formula

$$a_{s+1} = a_s \frac{s + l + 1 - A}{(s+1)(s+2l+2)}. \tag{265}$$

In order that the series reduce to a polynomial* (whose degree we shall indicate by n'), we must have $a_{n'+1} = 0$, $a_{n'} \neq 0$. Hence

$$n' + l + 1 - A = 0,$$

from which we get that A must be equal to the integer

$$n = n' + l + 1. \tag{266}$$

Finally, solving for $|E|$ from (259), we find that if E is negative, it must have one of the eigenvalues

$$E_n = -\frac{2\pi^2 m Z^2 e^4}{h^2 n^2}. \tag{267}$$

These are the same energy levels as given by the Bohr theory (see §16 of Part I), in perfect agreement with experience, as we have seen. The integer n corresponds to the principal quantum number, and the energy depends on it solely.

We note that with the aid of (267) the expression (256) for r_0 becomes

$$r_0 = na_1, \qquad \text{where } a_1 = \frac{h^2}{4\pi^2 m Z e^2}. \tag{268}$$

---

* If it does not, it can be shown (see e.g. No. 10c of the Bibliography, section 29b) that ω behaves asymptotically as e^x, and therefore y does not meet the required condition at infinity.

It is useful to mention that the constant a_1 defined in this manner is equal to the radius a_1 of the first of the circular orbits furnished by the Bohr theory [formula (16), page 38].

CASE 2: $E > 0$. In this case equation (258), for x tending toward ∞, approaches the form

$$\frac{d^2y}{dx^2} + \frac{1}{4}y = 0, \tag{269}$$

whose general solution is

$$c_1 e^{ix/2} + c_2 e^{-ix/2},$$

which is always finite for $x \to \infty$. Hence any integral of (258) will remain finite for $x \to \infty$. Therefore we are not forced to impose any limitations upon A; that is, any (positive) value of A is an eigenvalue of (258). Hence any (positive) value of E is an eigenvalue of (255) (*continuous spectrum of eigenvalues*). The fact that the energy may have any positive value is in perfect agreement with what is found in the Bohr-Sommerfeld theory. In fact, the positive values of E correspond to hyperbolic orbits, and hence neither to periodic, nor to multiply periodic motions, which are therefore not to be quantized.[5]

If we proceed as in the previous case, setting

$$y = e^{\pm ix/2}v(x), \tag{270}$$

we are led to the equation for v

$$v'' \pm iv' + \frac{A - l(l+1)}{x^2}v = 0, \tag{271}$$

and setting, as before,

$$v = x^{l+1}\omega, \qquad \omega = \sum_{s=0}^{\infty} a_s x^s, \tag{272}$$

we find for a_s the recursion formula

$$a_{s+1} = a_s \frac{-A \mp i(l+s+1)}{(2l+s+2)(s+1)}. \tag{273}$$

---

[5] We might even say that these states do not represent a true hydrogen atom but only the combination of an electron and a nucleus which, after approaching each other slightly, recede from each other indefinitely. Therefore, these solutions are of interest in collision theory.

As can be seen, since A is real, it is impossible for the numerator of this fraction to vanish, and therefore the series no longer reduces to a polynomial. The solution is a linear combination of the two contained in formula (270), namely,

$$y = e^{ix/2}x^{l+1}\sum_{s=0}^{\infty} a_s x^s + e^{-ix/2}x^{l+1}\sum_{s=0}^{\infty} a_s^* x^s, \tag{274}$$

in which the coefficients of the second sum are the complex conjugates of the coefficients in the first sum, so that y turns out real. It may easily be verified that these series converge absolutely for any value of x.

49. The eigenfunctions of hydrogenlike systems for E < 0. We shall now deal with the form of the eigenfunctions for the case where $E < 0$ (which is the most interesting one since it corresponds to a stable system). Since the factor depending on θ and φ (that is, the spherical harmonic Y_{lm}) was already discussed in §46, being common to all problems with central forces, there remains the factor $R(r)$ to be examined. Recalling the steps (249), (257), (260), and (262), we have

$$R(r) = \frac{y(r)}{r} = \frac{2}{r_0 x}\, e^{-x/2}v(x) = \frac{2}{r_0}\, e^{-x/2}x^l\omega(x),$$

where $\omega(x)$ is a polynomial of degree n' satisfying the differential equation (264), which we shall now write in the form

$$x\omega'' + (2l + 2 - x)\omega' + n'\omega = 0, \tag{264'}$$

and which is determined by the recursion relation (265), once the first coefficient a_0 has been fixed at will. The study of these polynomials is facilitated by the fact that they are related to a class of functions which have been studied for a long time and which possess important properties: the *Laguerre polynomials.*

Note on the Laguerre polynomials. The Laguerre polynomial of degree K, which is indicated by $L_K(x)$, is defined by the formula

$$L_K(x) = e^x \frac{d^K}{dx^K}\,(x^K e^{-x}). \tag{275}$$

Following are the explicit expressions of the first few Laguerre polynomials:

$$\begin{aligned} L_0 &= 1 \\ L_1 &= 1 - x \\ L_2 &= 2 - 4x + x^2 \\ L_3 &= 6 - 18x + 9x^2 - x^3 \\ &\cdots\cdots\cdots\cdots \end{aligned}$$

The interest of these equations lies in the fact that they are solutions of a well-known differential equation, as may be seen in the following manner. Letting

$$u(x) = x^K e^{-x}, \tag{276}$$

such that

$$L_K = e^x u^{(K)}, \tag{275'}$$

where by $u^{(K)}$ we denote, as we shall always do systematically, the Kth derivative of u, we have, upon a first differentiation of (276),

$$xu^{(1)} = (K - x)u.$$

We differentiate this relation another $(K + 1)$ times; for this operation it is useful to note that the known formula of Leibnitz for the nth derivative of a product, for the case in which the product is of the form $x\varphi(x)$, reduces to $\frac{d^n}{dx^n}(x\varphi) = x\varphi^{(n)} + n\varphi^{(n-1)}$. We then obtain for u the following differential equation:

$$xu^{(K+2)} + (1 + x)u^{(K+1)} + (K + 1)u^{(K)} = 0.$$

Substituting the $u^{(K)}$ obtained from (275′), we get the characteristic equation for the Kth Laguerre polynomial:

$$xL_K'' + (1 - x)L_K' + KL_K = 0. \tag{277}$$

It is to be noted that the Laguerre polynomials are not eigenfunctions of this equation, nor are they orthogonal. However, as may be shown[6] by successive integrations by part, utilizing (275), they have the property

$$\int_0^\infty L_K(x)L_{K'}(x)\, e^{-x}\, dx = \delta_{KK'}, \tag{278}$$

---

[6] For this and other properties of the Laguerre polynomials, see, for instance, No. 25 or No. 34 of the Bibliography.

which may be expressed as follows: the functions $e^{-x/2}L_K$ constitute an orthonormal system in the interval 0 to ∞.

Another important property of the Laguerre polynomials, which we shall merely mention here, is the one expressed by the recursion formula

$$L_{K+1} + (2K + 1 - x)L_K + K^2L_{K-1} = 0. \tag{279}$$

Associated Laguerre polynomials. By differentiating equation (277), we obtain

$$xL_K''' + (2 - x)L_K'' + (K - 1)L_K' = 0 \tag{280}$$

which may be considered as an equation of the second order in the function L_K'; that is, $L_K^{(1)}$. This one then satisfies the equation

$$xL_K^{(1)''} + (2 - x)L_K^{(1)'} + (K - 1)L_K^{(1)} = 0,$$

which is the analogue of (277), except that the second coefficient is increased by 1 and the third is decreased by 1.

To this equation we may again apply the same procedure. Differentiating j times, we find that the function $L_K^{(j)}$, that is, the jth derivative of L_K, satisfies the equation

$$xL_K^{(j)''} + (j + 1 - x)L_K^{(j)'} + (K - j)L_K^{(j)} = 0. \tag{281}$$

This function, evidently a polynomial of degree $(K - j)$, is sometimes called *associated Laguerre polynomial* and is defined by

$$L_K^{(j)} = \frac{d^j}{dx^j}\left[e^x \frac{d^K}{dx^K}(x^K e^{-x})\right]. \tag{282}$$

We note that for $j \geq K$ the polynomial $L_K^{(j)}$ is identically zero.

The associated polynomials correspond to a given superscript j and to different K-values, and when multiplied by $x^{j/2} e^{-x/2}$ give rise to functions which are orthogonal over the interval 0 to ∞. Specifically, we have

$$\int_0^\infty L_K^{(j)}(x)L_{K'}^{(j)}(x)\, x^j e^{-x} dx = \begin{cases} 0 & (\text{for } K \neq K') \\ \dfrac{(K!)^3}{(K-j)!} & (\text{for } K = K'). \end{cases} \tag{282'}$$

We now note that equation (264′), which ω satisfies, is identical with equation (281) of the associated Laguerre polynomials, provided that we take

$$j + 1 = 2l + 2, \qquad K - j = n';$$

that is,

$$j = 2l + 1, \qquad K = 2l + 1 + n' = n + l.$$

On the other hand, ω too is a polynomial of degree n' just like $L_{n+l}^{(2l+1)}$. Hence they can differ at most by a constant factor, which we shall designate by N'_{nl}. Therefore we write, placing in evidence the fact that ω and hence R depend on the indices n and l,

$$\omega_{nl} = \frac{1}{N'_{nl}} L_{n+l}^{(2l+1)}, \tag{283}$$

$$R_{nl} = \frac{2}{r_0 N'_{nl}} e^{-x/2} x^l L_{n+l}^{(2l+1)}(x). \tag{284}$$

The factor N'_{nl} is determined from the normalization condition (252). We find

$$N'^2_{nl} = r_0 \frac{n[(n+l)!]^3}{(n-l-1)!},$$

so that the explicit expression of R_{nl} as a function of r is the following, where we have put for r_0 the expression (268), to show more clearly its dependence upon n:

$$R_{nl}(r) = \sqrt{\frac{4(n-l-1)!}{a^3 n^4 [(n+l)!]^3}}\, e^{-\frac{r}{na}} \left(\frac{2r}{na}\right)^l L_{n+l}^{(2l+1)}\left(\frac{2r}{na}\right). \tag{285}$$

The explicit expressions corresponding to the first few values of n and l are, setting $\rho = r/a$,

$$R_{10} = \frac{2}{a^{3/2}} e^{-\rho},$$

$$R_{20} = \frac{1}{\sqrt{2}\, a^{3/2}} e^{-\rho/2}\left(1 - \frac{1}{2}\rho\right),$$

$$R_{21} = \frac{1}{2\sqrt{6}\, a^{3/2}} e^{-\rho/2} \rho,$$

$$R_{30} = \frac{2}{3\sqrt{3}\, a^{3/2}} e^{-\rho/3}\left(1 - \frac{2}{3}\rho + \frac{2}{27}\rho^2\right),$$

$$R_{31} = \frac{8}{27\sqrt{6}\, a^{3/2}} e^{-\rho/3}\rho\left(1 - \frac{1}{6}\rho\right),$$

$$R_{32} = \frac{4}{81\sqrt{30}\, a^{3/2}} e^{-\rho/3} \rho^2.$$

Equation (285), together with the expression already found for Y_{lm} [see formulas (245) and (246)], permits us to write down the complete expression for the eigenfunction corresponding to the quantum numbers n, l, m:

$$u_{nlm} = \sqrt{\frac{(l-|m|)!(n-l-1)!(2l+1)}{\pi a^3 n^4 \cdot (l+|m|)![(n+l)!]^3}} \cdot e^{im\varphi} \sin^{|m|}\theta \frac{d^{|m|}P_l(\cos\theta)}{d(\cos\theta)^{|m|}} e^{-\frac{r}{an}} \left(\frac{2r}{na}\right)^l L_{n+l}^{(2l+1)}\left(\frac{2r}{na}\right). \quad (286)$$

We observe that two eigenfunctions corresponding to values of m which are equal but of opposite sign differ only in the sign of the exponent $im\varphi$, and hence are complex conjugates.

It is interesting to get an intuitive picture of the distribution of the function uu^* around the nucleus—that is, of the "probability cloud" of the electron—as given by the preceding formulas.

First of all, we recall [see formula (248)] that in states in which $l = 0$ (*s-states*) u does not depend on θ or φ but only on r. Thus we have a spherically symmetrical cloud. In particular, the state corresponding to the lowest energy level, or ground state, is such a state. In fact, $n = 1$ necessarily implies $l = 0$ (and $m = 0$).

In that case, from the expression for R_{10} and from (248) we obtain

$$u = \frac{1}{\sqrt{\pi a^3}} e^{-r/a}$$

and hence

$$|u|^2 = \frac{1}{\pi a^3} e^{-2r/a}.$$

The density of the cloud decreases exponentially toward infinity. It is represented by Fig. 39*a*, in which the blackening at each point is approximately proportional to the probability density $|u|^2$ of finding an electron there. Hence we can say that the atom has no definite contour but, strictly speaking, extends over all space, since there exists a certain probability of finding the electron at any distance from the nucleus. However, as can be seen, this probability becomes rather small as soon as r reaches a value two or three times a (for example, the probability of finding the electron outside of the sphere of radius $3a$ is about 0.062). In that sense we may say that the dimensions of the atom in the ground state are of the order of a, that is, of the same order as predicted by the Bohr theory (radius of the first circular orbit).

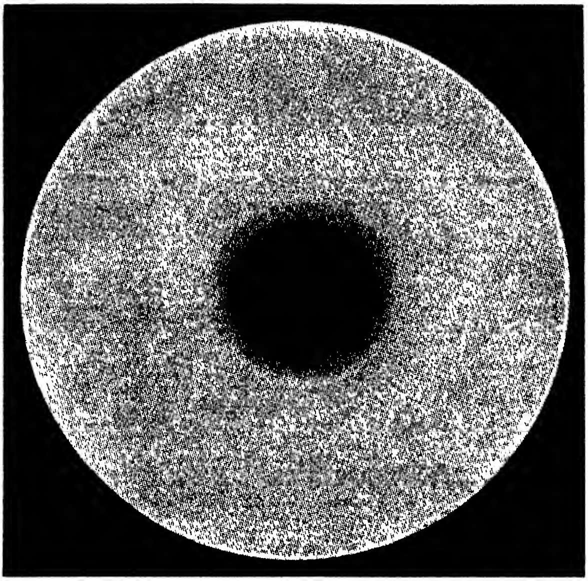

Fig. 39a

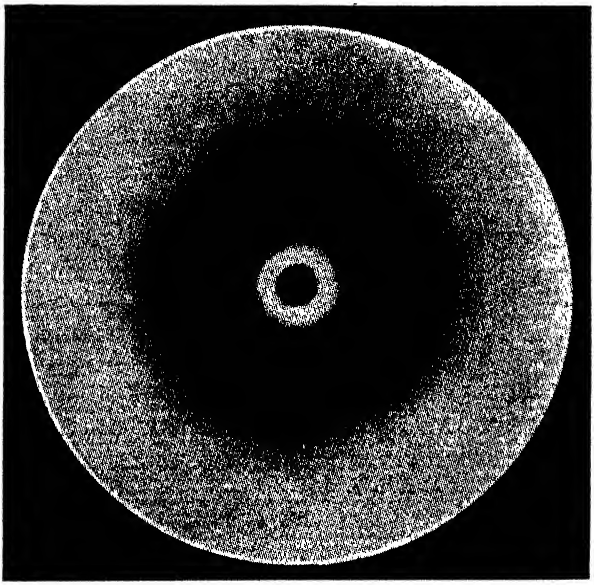

Fig. 39b

Analogous arguments may be made concerning the other states. Those for which $l \neq 0$ do not have spherical symmetry; however, even for these it is of interest to consider $R(r)$, or better $R^2(r)$ since, once we have fixed a straight line starting at the nucleus (that is, fixed θ and φ), the density along that line varies proportionally to R^2. In Fig. 40 the curve $R(r)$ is shown for the states $n = 1, n = 2, n = 3$. Also of interest is the function $z(r) = R^2(r) \cdot r^2$, whose meaning is the following: the probability of finding the electron within the infinitesimal spherical shell of radius r and thickness dr is evidently

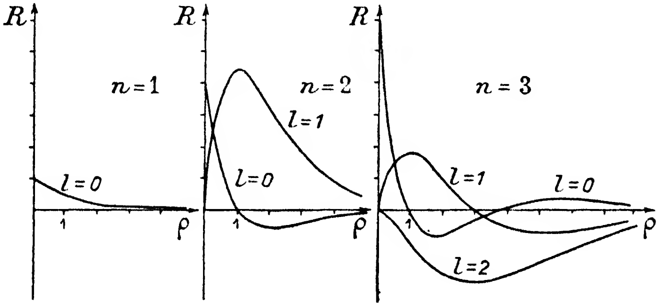

Fig. 40

$$dr \int\int |u|^2 r^2 \sin\theta \, d\theta \, d\varphi,$$

and because of (246′) this reduces to

$$R^2 r^2 \, dr = z(r) \, dr.$$

The function $z(r)$ is shown graphically in Fig. 41 for the same states as in Fig. 40.

As may be seen from Figs. 40 and 41, there generally exist values of r for which $R(r)$ vanishes. These correspond to spheres on which the probability density is zero (*nodal spheres*). Figure 39b, representing the "probability cloud" for hydrogen in the state $n = 2$, $l = 0$ (2s-state) shows an example of this condition. As may be seen from expression (285) for R, this function vanishes for $r = 0$ (except for the case $l = 0$), for $r = \infty$, and also for the roots of the polynomial $L_{n+l}^{(2l+1)}$. The spheres of radius zero and radius infinity are not counted as nodal spheres, so that the number of these spheres is equal to the number of positive roots of the polynomial $L_{n+l}^{(2l+1)}$, and it may be shown (*Perron's theorem*) that all the n' roots of this polynomial of degree n' are positive, so that *the number of nodal spheres is equal to the radial quantum number* $n' = n - l - 1$.

50. Selection rules. As has been mentioned in §32, the intensity and state of polarization of the emitted radiation in the quantum jump from a state n to a state m are determined by the quantities x_{nm}, y_{nm}, z_{nm} defined by (144); that is, they are obtained by integration over all space of the product of the eigenfunctions corresponding to the two states in question. We shall rapidly outline[7] the calculation of these quantities for the case of an electron under the

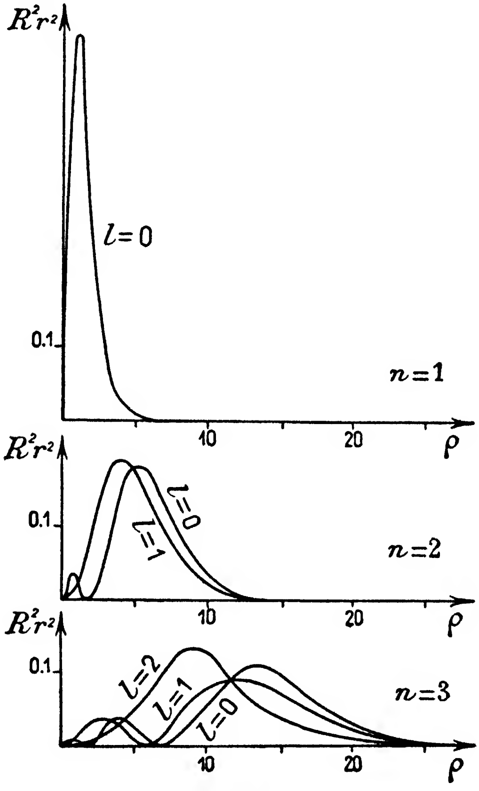

Fig. 41

---

[7] For a more complete treatment see, for example, No. 1 of the Bibliography.

action of a central force (hence, in particular, for hydrogenlike systems).

Since each state is characterized by three quantum numbers, we must substitute a triplet of indices for each of the indices n and m just used. We shall denote the triplets by n_1, l_1, m_1 and n_2, l_2, m_2, respectively (and sometimes only by 1 and 2), so that the quantities to be calculated will be $x_{n_1l_1m_1n_2l_2m_2}$ (which we shall write x_{12} for short), and so forth.

It is convenient to introduce, instead of the Cartesian coordinates x and y, their linear combinations

$$\xi = x + iy, \qquad \eta = x - iy,$$

so that, instead of x_{12}, y_{12}, z_{12}, we shall calculate

$$\left.\begin{aligned} \xi_{12} &= \int_S \xi u_1^* u_2 \, dS, \\ \eta_{12} &= \int_S \eta u_1^* u_2 \, dS, \\ z_{12} &= \int_S z u_1^* u_2 \, dS. \end{aligned}\right\} \tag{287}$$

It is evident that one may pass from ξ_{12} and η_{12} to x_{12} and y_{12} by means of the relations

$$\xi_{12} = x_{12} + iy_{12}, \qquad \eta_{12} = x_{12} - iy_{12}.$$

Introducing the polar coordinates r, θ, φ, we evidently have

$$\xi = r \sin\theta \, e^{i\varphi}, \qquad \eta = r \sin\theta \, e^{-i\varphi}, \qquad z = r\cos\theta,$$
$$dS = r^2 \sin\theta \, d\theta \, d\varphi \, dr.$$

Keeping in mind that each of the functions u_1 and u_2 is made up of the product of the three factors $\Phi(\varphi)$, $\Theta(\theta)$, $R(r)$, we see that each of the triple integrals (287) splits into the product of three simple integrals:

$$\left.\begin{aligned} \xi_{12} &= \int_0^{2\pi} e^{i\varphi}\Phi_1^*\Phi_2 \, d\varphi \int_0^{\pi} \Theta_1^*\Theta_2 \sin^2\theta \, d\theta \int_0^{\infty} r^3 R_1^* R_2 \, dr \\ \eta_{12} &= \int_0^{2\pi} e^{-i\varphi}\Phi_1^*\Phi_2 \, d\varphi \int_0^{\pi} \Theta_1^*\Theta_2 \sin^2\theta \, d\theta \int_0^{\infty} r^3 R_1^* R_2 \, dr \\ z_{12} &= \int_0^{2\pi} \Phi_1^*\Phi_2 \, d\varphi \int_0^{\pi} \Theta_1^*\Theta_2 \cos\theta \sin\theta \, d\theta \int_0^{\infty} r^3 R_1^* R_2 \, dr. \end{aligned}\right\} \tag{288}$$

Let us first deal with the integrals with respect to φ. The first of these becomes, upon substitution of the expressions for Φ_1 and Φ_2 according to (229′),

$$\frac{1}{2\pi}\int_0^{2\pi} e^{i(1-m_1+m_2)\varphi}\,d\varphi.$$

Thus it is equal to zero except for the case where the exponent vanishes, that is, when $m_2 = m_1 - 1$, in which case the integral is equal to 1. Similarly, the second integral with respect to φ vanishes except when $m_2 = m_1 + 1$, and the third integral vanishes unless $m_2 = m_1$. Hence, indicating by Δm the difference $m_2 - m_1$ (change in magnetic quantum number), we may say that in order that at least one of the three integrals not be zero, we must have

$$\Delta m = \pm 1, 0. \tag{289}$$

Otherwise all three quantities ξ_{12}, η_{12}, z_{12} vanish, and hence also x_{12}, y_{12}, z_{12}, which means (see §32) that the corresponding spectral line has intensity zero, or else that a quantum jump in which m does not change by ±1 or zero is "forbidden." Hence we have in (289) the selection rule for the magnetic quantum number. It may also be added that to a transition $\Delta m = 0$ (only z_{12} different from zero) there corresponds the emission of plane-polarized radiation like that which would be emitted by an oscillator vibrating along the z-axis. To the transitions $\Delta m = \pm 1$ there correspond $z_{12} = 0$ and $x_{12} = \mp i y_{12}$; that is, the radiation corresponds to that emitted by an electric moment rotating in the xy-plane (in the clockwise or counterclockwise direction) and hence is circularly polarized when viewed along the z-direction and plane-polarized when observed in the xy-plane.

In analogous fashion we obtain the selection rule for the azimuthal quantum number l from a consideration of the integrals with respect to θ. Since the calculation is somewhat lengthy, we shall confine ourselves to mentioning the main result. We find that the three integrals (in which, of course, m_2 must be replaced by $m_1 + 1$, $m_1 - 1$, and m_1, respectively) all vanish except for the case where $l_2 = l_1 \pm 1$, so that for a transition $\Delta l = l_2 - l_1$ we find the selection rule

$$\Delta l = \pm 1. \tag{290}$$

Finally, a consideration of the integral with respect to r (which is the same in all three formulas) does not furnish any selection rule, since it does not vanish (once the selection rules for m and l are satisfied). Therefore the jump Δn of the principal quantum number n may be anything.

Actually, experiments show that in spectra generally only the lines satisfying the selection rules just found for m and l occur. These rules, indeed, already had been discovered empirically before the rise of wave mechanics[8] and were justified in various ways in the Bohr-Sommerfeld theory (for example, by means of the *correspondence principle;* see §64).

The actual calculation of the integrals contained in (288) permits the quantitative evaluation of the probability for the quantum jump and the intensity of the corresponding spectral line. In the not very numerous cases in which it has been possible to make a comparison with experiment, the results of the theory have been confirmed.

---

[8] In certain exceptional cases, lines which would be "forbidden" by the selection rules also show up, though with considerably reduced intensity. This effect depends on various factors which the theory accounts for perfectly in a further approximation (for instance, by considering weak *quadrupole* transitions in addition to the strong *dipole* transitions which yield the selection rules). It is also to be understood that the transition between two quantum states, even though prohibited by a selection rule, may always occur by an indirect route (that is, through intermediate levels) or by a process different from radiation, such as by collision for example.

CHAPTER 9

The Bohr-Sommerfeld Theory

51. The method of Wentzel and Brillouin. In this chapter we shall first of all treat an important method, originated by Wentzel[1] and Brillouin[2] and perfected by Kramers[3] and many others, of finding approximate expressions for the eigenfunctions and eigenvalues of the Schrödinger equation. From this method we shall obtain a *quantization rule* which essentially coincides with the one postulated by Sommerfeld, upon which atomic mechanics was founded until 1925 (see Part I). Having thus obtained the fundamentals of the Bohr-Sommerfeld theory as a first approximation of wave mechanics, we shall treat some of the most important results of that theory, which it would be too long or too intricate to deduce in a rigorous manner from the Schrödinger equation.

We shall now expose the method of Wentzel and Brillouin as applied to the one-dimensional case, to which the more general cases may be reduced by means of the separation of variables. Hence we shall write the one-dimensional Schrödinger equation for a state of energy E [(146), §34] in the form

$$u'' + \frac{4\pi^2}{h^2} p^2 u = 0, \tag{291}$$

where we have put

$$p(x) = \sqrt{2m(E - U)}. \tag{292}$$

We note that for all values of x for which $U < E$, this function p is identical with the classical expression for the momentum (taken in absolute value) of the particle at the point x. In the regions where $U > E$, p is imaginary—an indication, according to classical mechanics, that these regions of the x-axis are inaccessible to the particle. We shall suppose that the potential energy U has a shape

---

[1] *Zeits. f. Physik* **38**, 518 (1926).

[2] *Comptes Rendus* **183**, 24 (1926); *J. de Phys.* **7**, 353 (1926).

[3] *Zeits. f. Physik* **39**, 828 (1926).

of the type shown in Fig. 42, so that there is only one region AB (region II, from x_1 to x_2) in which p is real, whereas it is imaginary with positive coefficient in the other two regions (I and III). Classically, the particle would execute oscillations between A and B.

The linear second-order equation (291) may be transformed into one of the first order which is nonlinear (of the Riccati type) by means of the well-known transformation from the theory of differential equations,

$$u = e^{\frac{2\pi i}{h}\int^{x} y\,dx} \tag{293}$$

(where the lower limit of the integral is any fixed value of x), that is, by taking as a new unknown function

$$y = \frac{h}{2\pi i}\cdot\frac{u'}{u}.$$

In fact it may immediately be verified, upon substituting into (291), that y must satisfy the Riccati equation

$$\frac{h}{2\pi i}\,y' = p^2 - y^2. \tag{294}$$

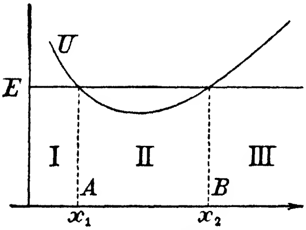

Fig. 42

If we consider that, in general, the results of wave mechanics approach those of ordinary mechanics if we neglect quantities of the order of h, we are led to believe that we may obtain a first approximation by neglecting h, a second approximation by neglecting h^2 and all higher powers, and so forth. This procedure *suggests* trying to find for y an approximate expression Y, of the form

$$Y = Y_0 + \frac{h}{2\pi i}\,Y_1 + \left(\frac{h}{2\pi i}\right)^2 Y_2 + \cdots + \left(\frac{h}{2\pi i}\right)^k Y_k, \tag{295}$$

where Y_0, Y_1, . . . are functions of x which may be determined formally by substituting Y for y in (294) and equating the coefficients of each power of $h/2\pi i$ on both sides. We then find immediately for the first few terms the following recursion relations:

$$Y_0 = \pm p, \tag{296}$$

$$Y_1 = -\frac{Y_0'}{2Y_0}, \tag{296'}$$

$$Y_2 = -\frac{Y_1' + Y_1^2}{2Y_0}, \tag{296''}$$

and the successive terms may readily be found.

As can be seen, we obtain for Y two expressions (which we shall indicate by Y_a and Y_b), depending on whether we take $Y_0 = +p$ or $Y_0 = -p$ in the first term. One of these is obtained from the other by changing the sign of all odd terms. They are (writing only the first two terms, and representing by dots the succeeding ones up to the kth term)

$$Y_a = p - \frac{h}{2\pi i}\frac{p'}{2p} + \cdots, \tag{297}$$

$$Y_b = -p - \frac{h}{2\pi i}\frac{p'}{2p} + \cdots. \tag{297'}$$

These two expressions approximately represent two different integrals of (294).

It is useful here to clarify the significance of these approximate solutions. If, rather than breaking off one of these series at the kth term, we continue it indefinitely, we obtain a *diverging* series, and hence we may not consider y approximated by a *series* of the type (297) or (297′). However, we may show that in general the first few terms decrease rapidly and that, if the sum is limited to these terms, Y_a and Y_b represent, to a good approximation, two particular integrals of (294) for all points of the x-axis, *except* those near the two critical points A and B where $p = 0$. In fact, at these points Y_a and Y_b become infinite, whereas the solutions of (294) should not possess any singularities.

From each of the two functions Y found in this way we may obtain an approximate solution of (291) by means of (293), and hence any solution of the latter may be approximated by an expression of the type

$$u = c_a e^{\frac{2\pi i}{h}\int^x Y_a\,dx} + c_b e^{\frac{2\pi i}{h}\int^x Y_b\,dx}, \tag{298}$$

c_a and c_b being constants. This approximation, however, ceases to be valid in the vicinity of the two critical points A and B. It follows that in order to represent the same solution u in the three regions I, II, III of the real axis, different values must be given to the constants c_a and c_b in each of the regions. We shall indicate these values by c_a', c_b', c_a'', c_b'', c_a''', c_b''' respectively. We shall now see how these values must be related to each other, confining ourselves from now on to the approximation given by the first two terms of Y_a and Y_b, so that (298) may also be written [see (297) and (297′)], fixing the lower limit at x_1 and performing the integration along the real axis by convention,

$$u = \frac{c_a}{\sqrt{p}} e^{\frac{2\pi i}{h}\int_{x_1}^{x} p\,dx} + \frac{c_b}{\sqrt{p}} e^{-\frac{2\pi i}{h}\int_{x_1}^{x} p\,dx}. \tag{299}$$

We note first of all that for $x \rightarrow -\infty$, U tends toward $+\infty$ and hence p toward $+i\infty$. It follows that the exponent of the first term tends toward $+\infty$; and therefore, in order that u may tend toward zero for $x \rightarrow -\infty$, as it should, we must have $c_a' = 0$. Thus in region I, u is reduced to

$$u = \frac{c_b'}{\sqrt{p}} e^{-\frac{2\pi i}{h}\int_{x_1}^{x} p\,dx}. \tag{300}$$

In region II, (299) may also be written (setting $c_a'' = C'' e^{i\delta}$, $c_b'' = C'' e^{-i\delta}$)

$$u = \frac{2c''}{\sqrt{p}} \cos\left(\frac{2\pi}{h}\int_{x_1}^{x} p\,dx + \delta\right). \tag{301}$$

In order to relate these two expressions, it is necessary to find an approximate expression for u that is valid in the zone where the preceding formulas break down, namely, in the vicinity of the point A. This expression may be obtained by replacing the potential curve at that point by a short straight-line segment (that is, by considering the force field to be uniform over the short portion considered), or by putting the following into the Schrödinger equation:

$$\frac{8\pi^2 m}{h^2}(E - U) = K_1(x - x_1),$$

where K_1 is a constant (> 0). The Schrödinger equation may then be solved rigorously, and u is found to be expressed by a Bessel function or, in another form, by means of a definite integral. This solution (which has no singularity for $x = x_1$) must be continuously joined, on one side (for $x < x_1$) to (300), on the other side to (301); this determines the values of the constants C'' and δ of (301). The calculations will be omitted, but it is found[4]

---

[4] See Kramers, *loc. cit.*, or No. 22 of the Bibliography. We may also substitute a section of a step for the potential curve, in the vicinity of A. It may then be joined more easily to (301′), since this method avoids the use of Bessel functions [(see E. Persico, *Nuovo Cimento* **XV**, 133 (1938)].

that in region II the continuation of the solution (300) is represented by

$$u = \frac{2c_b'}{\sqrt{ip}} \cos\left(\frac{2\pi}{h}\int_{x_1}^{x} p\,dx - \frac{\pi}{4}\right). \tag{301'}$$

Proceeding to a consideration of region III, we find, by a method analogous to the preceding, that for u to vanish for $x \to +\infty$, the second term of (299) must be absent; that is, u must have the form

$$u = \frac{c_a'''}{\sqrt{p}} e^{\frac{2\pi i}{h}\int_{x_2}^{x} p\,dx},$$

where c_a''' is a constant, undetermined for the present and of no interest here. This expression must be joined to one of the form (301), valid in region II, across the critical point B. The junction may be made by the same procedure followed for the point A, and we find that u must be represented, in region II, by

$$u = \frac{2c_a'''}{\sqrt{ip}} \cos\left(\frac{2\pi}{h}\int_{x}^{x_2} p\,dx - \frac{\pi}{4}\right). \tag{301''}$$

Now in order that this expression be identical with (301′), the two cosine terms must be equal, in absolute value, for any x. This condition requires that the sum or the difference of their arguments be an integral multiple of π. But since the difference of the arguments depends upon x, whereas their sum turns out to be constant, as we shall see presently, it will be the sum which must be equated to $n\pi$ (with n an integer), thus yielding

$$2\int_{x_1}^{x_2} p\,dx = \left(n + \frac{1}{2}\right)h. \tag{302}$$

Recalling that according to classical mechanics, the particle would oscillate between x_1 and x_2 with momentum $+p$ in the motion from x_1 to x_2, and $-p$ in the other half cycle, we may interpret the left-hand member as the integral of $p\,dx$ (where p now indicates the momentum taken algebraically) extended over a complete oscillation of the classical motion, which is denoted by the symbol $\oint$. Therefore the last formula may be written

$$\oint p\,dx = \left(n + \frac{1}{2}\right)h. \tag{303}$$

This formula does not contain the eigenfunctions, but only E (contained in p). It may therefore serve to determine (approximately) the various eigenvalues of the problem without going through the Schrödinger equation. Upon assigning to n successive integral values, we obtain the successive eigenvalues of E. Equation (303) thus represents an approximate *quantization rule* valid for all cases in which the potential has the behavior of Fig. 42.

It is important to mention that if n is rather large, (303) may practically be replaced by

$$\oint p\,dx = nh. \tag{303'}$$

Let us now consider the case where x represents an angle rather than a Cartesian coordinate, which may vary from 0 to 2π (for example, the angle θ in plane-polar coordinates). In this case, whose importance will be better appreciated in what follows, U (and hence p) is a periodic function of x with period 2π, and u, rather than approaching zero at infinity, must also be periodic in x with period 2π. Let us suppose that E is sufficiently large for p to be real everywhere; that is, we refer to the case in which the classical motion would be rotational (not oscillatory). Then u must have the form (301) everywhere (x_1 is any fixed value of x) and will be periodic with period 2π only if

$$\frac{2\pi}{h}\int_{x_1}^{x_1+2\pi} p\,dx = 2\pi n,$$

or if

$$\oint p\,dx = nh,$$

with $\oint$ indicating the integral extended over a period. Hence the quantization condition is in this case exactly (303′) rather than (303).

52. The Sommerfeld conditions. The considerations of the preceding section justify the very often spectacular success of the method of quantization postulated by Sommerfeld long before the rise of wave mechanics. This method consists, as we have mentioned in Part I, in treating the atom at first as a system of material points subject to the ordinary laws of classical mechanics (or, in a further approximation, of relativistic mechanics), and then in adding to these laws some restrictive conditions (*Sommerfeld conditions*) which allow only certain motions among the infinite number of motions permitted by ordinary mechanics, and hence only certain values of the energy. Before formulating the Sommerfeld conditions, we must recall a few fundamental concepts of rational mechanics.

Let us consider a mechanical system with f degrees of freedom, with constraints independent of time, and subject to conservative forces. We shall refer it to a system of *generalized coordinates* $q_1, q_2, \ldots q_f$ (which, in particular, may be the Cartesian coordinates of the single particles, if we are dealing with a system of points) to which there correspond as many *momenta*[5] $p_1, p_2, \ldots p_f$; and let us suppose that its motion is such that each of the q_i and the corresponding *conjugate* momenta p_i are periodic[6] functions of the time with period T_i. Motions which—by a judicious choice of coordinates—exhibit this property, are called *multiple periodic* motions. Incidentally, we note that in general, all the f periods T_i will be different. If these periods are equal or commensurable, the motion is *periodic* because, after a time which is a common multiple of all the T_i, the motion repeats identically. However, in general this condition does not obtain; hence the periodic motions are special cases of multiple periodic motions, characterized by the fact that we can find f *integers* $m_1, m_2, \ldots m_f$ such that

$$\frac{m_1}{T_1} = \frac{m_2}{T_2} = \cdots = \frac{m_f}{T_f},$$

---

[5] As is known from mechanics, the kinetic energy T of the system is a function of the q and the $\dot{q}$, and we call *conjugate momenta* the quantities

$$p_i = \frac{\partial T(q, \dot{q})}{\partial \dot{q}_i} \quad (i = 1, 2, \ldots f).$$

Since T is a quadratic function of the $\dot{q}$, the momenta will be linear functions of the $\dot{q}$; it is possible to solve for them and express the $\dot{q}$ as linear functions of the p.

In particular, if the coordinates q are the ordinary Cartesian coordinates x, y, z of a point, the corresponding momenta are the components of the linear momentum; that is, since $T = \frac{1}{2}m(\dot{x}^2 + \dot{y}^2 + \dot{z}^2)$,

$$p_x = m\dot{x}, \qquad p_y = m\dot{y}, \qquad p_z = m\dot{z}.$$

[6] If one of the q represents an angle such that upon increasing it by 2π we again obtain the same configuration of the system (for example, the angle in a system of plane-polar coordinates), we consider as period T_i relative to this coordinate the time required for the latter to increase by 2π. We then say that there is a degree of freedom of "rotation," while those coordinates which oscillate periodically between two limits are said to correspond to degrees of "libration" or of "oscillation."

or else, introducing the frequencies $\nu_i = 1/T_i$,

$$m_1\nu_1 = m_2\nu_2 = \cdot \cdot \cdot = m_f\nu_f.$$

For the present we shall exclude not only this case but also the more general case in which between the frequencies ν_i there exist one or more relations of the type

$$m_1\nu_1 + m_2\nu_2 + \cdot \cdot \cdot + m_f\nu_f = 0, \tag{304}$$

with integral coefficients. When there are g relations of this kind, the system is said to be *g-fold degenerate.* If the motion of a system with f degrees of freedom is periodic, it is therefore $(f - 1)$-fold degenerate. Degenerate systems occur quite often in questions of atomic physics, but we shall deal with them elsewhere.

Among the multiple periodic systems, we shall be particularly concerned with those possessing a system of generalized coordinates such that each of the momenta p_i is a function only of q_i and not of the others. This condition is to be expressed by saying that "the variables are separable." In fact, this statement is equivalent to saying that the Hamilton-Jacobi equation[7] can be satisfied by a function of the form

$$W = \sum_i f_i(q_i) \tag{308}$$

rather than by the more general form $W = f(q_1 \ . \ . \ . \ q_f)$. All these conditions seem very restrictive, but in practice the majority of systems which occur in the study of atomic mechanics satisfy them, and hence the Sommerfeld conditions which we are about to state may be applied to them.

---

[7] From analytic mechanics we recall the method of Jacobi for the integration of the equations of motion, when the constraints are independent of time and the forces are conservative. We express the total energy (sum of kinetic energy T and potential energy U) as a function of the q and p [the function $H(q, p)$ defined in this way is called the *Hamiltonian* of the system, and contains within itself everything needed to characterize the mechanical properties of the system]. Then we substitute for each p_i in the expression for H the quantity $\partial W/\partial q_i$ (where W is an unknown function of the q), and write the partial differential equation (*Hamilton-Jacobi equation*):

$$H\left(q, \frac{\partial W}{\partial q}\right) = \alpha_1, \tag{305}$$

where α_1 is an arbitrary constant. We are now to find a solution $W(q, \alpha_1, \alpha_2, \ . \ . \ . \ \alpha_f)$ of this equation which, in addition to α_1, contains $f - 1$ other arbi-

The motion of the system with f degrees of freedom depends, as we know, on $2f$ constants $\alpha_1, \alpha_2 \ldots \alpha_f, \beta_1, \beta_2 \ldots \beta_f$ which are defined by the initial conditions. Sommerfeld imposes a restriction upon these constants, forcing them to satisfy the f following conditions. For each coordinate, we calculate the so-called *phase integral*

$$J_i = \oint p_i \, dq_i, \tag{309}$$

extended over an entire cycle of the same coordinate. Since p_i depends only on q_i (by hypothesis) and on the f constants α, this integral will be a function $J_i(\alpha_1 \ldots \alpha_f)$ of the constants α only. We may also invert these functions and solve for the α as functions of the f constants J. The Sommerfeld conditions consist in setting each one of the integrals J equal to an integral multiple of Planck's constant, that is, in writing

$$\oint p_i \, dq_i = n_i h \quad (i = 1, 2, \ldots f), \tag{310}$$

where n_i is a nonnegative integer.[8] Hence there are as many conditions as there are degrees of freedom, and we introduce as many

---

trary constants $\alpha_2, \alpha_3 \ldots \alpha_f$. Having found W, we obtain the equations of motion by writing the following $2f$ relations between the q, the p, and t (from which the q and the p may be found explicitly as a function of t):

$$\frac{\partial W}{\partial \alpha_1} = t + \beta_1 \tag{306}$$

$$\frac{\partial W}{\partial \alpha} = \beta_j \qquad (j = 2, 3, \ldots f), \tag{306'}$$

$$p_i = \frac{\partial W}{\partial q_i} \qquad (i = 1, 2, 3, \ldots f), \tag{307}$$

where the β are f other arbitrary constants.

Of these equations, the (306′), which do not contain t, determine the form of the trajectory of each point; (306) determines the law according to which points move along the trajectory; and (307) determine the momenta and hence the velocities (as functions of the q and α). The constant α_1 has the physical meaning of *total energy*. The constants β which do not occur in (307) have the meaning of *phases*.

Often equation (305) may be solved by separation of variables, that is, by taking W of the form (308), where of course each f_i depends also on the f constants α. Then evidently (307) will assume the form $p_i = f_1'(q_i, \alpha)$, or each p_i depends solely on its conjugate q_i and on the α. This is the case to which the Sommerfeld conditions may be applied.

[8] In fact, it may be easily shown that the integral in the left member is never negative.

quantum numbers n_i which may replace the f constants α. Thus instead of introducing f constants capable of assuming continuous values, we introduce an equal number which take on discrete values.

Now since, as we have seen, one of the α is identified with the energy E of the system, we may regard the energy as a function of the f constants J, and hence of the n_i. We conclude that the energy may take on only discrete values, depending on the f integers $n_1, n_2, \ldots n_f$; these values may be designated by $E_{n_1 n_2 \cdots n_f}$.

Thus, as will be seen, the energy levels are generally found to be in good agreement with experimental values, and sometimes in excellent agreement with them.

In the special case of a single degree of freedom, the Sommerfeld condition coincides with (303′), which we have deduced in an approximate way from wave mechanics. However, we note that in the case of one oscillatory degree of freedom we have found that the best approximation is obtained by equating the integral to $(n + \frac{1}{2})h$, whereas in the Sommerfeld method it is equated to nh as for the rotational degrees of freedom. Actually, in these cases the introduction of half-integral numbers (that is, of the type $n + \frac{1}{2}$) instead of integral quantum numbers, generally improves the approximation given by Sommerfeld, as was also observed empirically even before the rise of wave mechanics; and in certain cases, such as that of the oscillator, it immediately furnishes the exact result. On the other hand, for rotational degrees of freedom the quantum numbers must be integers, as we have seen.

In the case of a system with several degrees of freedom and with separable variables, the Sommerfeld conditions may be found again in an analogous manner (through extension of the considerations of §51 by means of separation of variables), as a consequence of first approximation of the Schrödinger equation and of the conditions of regularity imposed upon the eigenfunctions.

53. Note on the choice of the coordinate system. Since a mechanical system may be referred to an infinite number of generalized coordinates, the following question arises: If instead of the system of the q, we adopt another coordinate system $q'_1, q'_2, \ldots q'_f$ (whose conjugate momenta will be called $p'_1, p'_2, \ldots p'_f$) and if we apply to them the Sommerfeld conditions

$$\oint p'_i \, dq'_i = n_i h, \qquad (311)$$

do we get the same quantum orbits? First it is necessary to observe that in order to be able to apply the conditions (311), we must be able to separate variables in the system of the q' as well. This statement implies (if the system is not degenerate) that the transformation leading from q to q' must be one of a special, limited class, namely, one which neither alters the coordinate lines nor modifies the integrals J (or, at the most, performs linear transformations on them, with integral coefficients and with a determinant of unity). In this case equations (311) determine the same trajectories as (310), generally yielding different values for n' in the two cases.

If the system is degenerate, the situation is different. Then the separation of variables is possible even for essentially different coordinate systems, which lead to values for J that are not reducible to the previous values by linear transformations with integral coefficients and unit determinant; and hence quantum trajectories are obtained which are different according to the system of reference selected. However, in these cases it also happens that the energy levels prove to be independent of the choice of reference system. This result is attributable to the fact that the energy levels have physical significance, whereas the orbits of the Bohr-Sommerfeld theory are only an analytic device (see §57).

Thus in the treatment of degenerate systems there is a certain degree of arbitrariness in the selection of the coordinate system. Consequently, we generally let ourselves be guided by the following criterion. We observe that if we *perturb* the system ever so slightly (for example, by slightly changing the forces which act upon it), the system generally ceases to be degenerate, because its periods T_i vary slightly and no longer satisfy relations (304). Now very often a system appears to be degenerate only because, for simplicity, we neglect some of the slight perturbations which are really present. For instance, an atom is usually considered uninfluenced by all external causes, but in reality it is always located in a magnetic field, no matter how small, caused by neighboring atoms, terrestrial magnetism, or other factors. It is therefore sufficient to take these perturbations into account and to look for a system of reference in which the variables are separable. This same system of coordinates is then chosen to treat the problem, even when the perturbations are being neglected. In fact, it is clear that if the perturbing forces are being reduced to zero gradually, their effect upon the motion will also tend to zero, whereas their effect in fixing a system of reference in which the Sommerfeld conditions apply is independent of their magnitude.

54. The harmonic oscillator. We shall again take up the problem studied in §39 in order to treat it by the Sommerfeld method. First, from mechanics we know that the particle would execute oscillations according to

$$x = A \sin (2\pi\nu_0 t - \varphi), \tag{312}$$

where A and φ are two arbitrary constants. The momentum conjugate to x is

$$p = m\dot{x} = 2\pi\nu_0 mA \cos (2\pi\nu_0 t - \varphi).$$

The total energy is [see (183) in §39]

$$E = U + \tfrac{1}{2}m\dot{x}^2 = 2\pi^2 m\nu_0^2 x^2 + \tfrac{1}{2}m\dot{x}^2 = 2\pi^2\nu_0^2 A^2 m$$

and is seen to contain only one of the two constants of integration. We shall now calculate the phase integral

$$J = \oint p\, dx = \oint m\dot{x}\, dx = m \int_0^{1/\nu_0} \dot{x}^2\, dt$$
$$= m4\pi^2\nu_0^2 A^2 \int_0^{1/\nu_0} \cos^2 (2\pi\nu_0 t - \varphi)\, dt.$$

Since the last integral is equal to $1/2\nu_0$, we have

$$J = 2\pi^2\nu_0 A^2 m = \frac{E}{\nu_0};$$

hence the Sommerfeld condition $J = nh$ gives for E the values

$$E_n = nh\nu_0 \quad (n = 0, 1, 2, \ldots). \tag{313}$$

Thus we find the law (already postulated by Planck in the theory of black-body radiation) that the energy of the oscillator is always an integral multiple of the "quantum" $h\nu_0$. If we use the formula (303) instead, which is a better approximation when dealing with a motion of the oscillatory type, we find

$$E_n = (n + \tfrac{1}{2})h\nu_0, \tag{313'}$$

that is, there is added, to the n quanta $h\nu_0$, a fixed amount of energy $\tfrac{1}{2}h\nu_0$ which the oscillator always possesses ("*zero-point energy*"; see also page 183). The same result may be obtained by the Schrödinger method and is confirmed in various ways by experience. Thus in this special case the method of Wentzel and

Brillouin (accidentally) furnishes the exact result, as far as the energy is concerned.

55. The rigid rotator and the Bohr atom. A *rigid rotator* is a system constituted of a rigid body revolving around a fixed axis, not subject to forces (or subject to forces whose moment is zero with respect to the axis). The system has a single degree of freedom; its position at each instant may be specified by means of an angular coordinate θ, expressing the angle between two planes passing through the axis of rotation, one being fixed in space and the other fixed in the body. The energy of the system, which reduces to kinetic energy alone, is given by

$$E = \tfrac{1}{2} I \dot{\theta}^2,$$

where I indicates the moment of inertia.

The momentum conjugate to the coordinate θ is hence (see footnote 5 in §52)

$$M = I\dot{\theta};$$

that is, it represents the *angular momentum.* As is well known, the motion takes place according to the law $M = \text{const.}$, so that

$$\oint M\, d\theta = M \oint d\theta = 2\pi M.$$

The Sommerfeld condition therefore gives

$$M = n \frac{h}{2\pi}, \tag{314}$$

and may be expressed by saying that *the angular momentum must be a multiple of* $h/2\pi$.[9] Hence the angular velocity must have one of the discrete values $n(h/2\pi I)$, and, substituting into the expression for E, we find for the energy the values

$$E_n = n^2 \frac{h^2}{8\pi^2 I}. \tag{315}$$

By chance, this result also coincides perfectly with the one obtained from a rigorous solution of the Schrödinger equation when generalized to include such a system.

---

[9] Since in many other cases too, the angular momenta result as multiples of $(h/2\pi)$, or at least as simple ratios to this quantity, $h/2\pi$ is often taken as unit of measurement for angular momentum (*quantum unit*). For example, the result expressed by (314) may be stated by saying, "The angular momentum of the rotator is given by an integral number n" (implying quantum units).

We note that the hydrogen atom as conceived by Bohr (see §16 of Part I), in which the electron is constrained to move in a circular orbit, may be mechanically identified with a rigid rotator, and hence (314) must also hold for a circular orbit. Thus we see that the condition postulated by Bohr to characterize the quantum orbits is contained as a special case within the Sommerfeld conditions.

56. Hydrogenlike systems. We shall now apply the Sommerfeld method to the hydrogen atom and, in general, to hydrogenlike systems, *without* the purely artificial restriction of circular orbits as adopted by Bohr in the first quantum theory of these systems and outlined in §16 of Part I. However, we shall still suppose that the nucleus is fixed, for the time being. Hence the system is one with three degrees of freedom.

(a) *Mechanical part.* It is known from rational mechanics that the motion of an electron under the action of a central attractive force of intensity $Z\,e^2/r$ (that is, analogous to the Newtonian attraction) takes place in elliptic orbits according to Kepler's laws, excluding the parabolic and hyperbolic orbits which correspond to states in which the electron is not bound to the nucleus. If polar coordinates r and ω are introduced in the plane of the orbit (with the pole at the nucleus and the polar axis directed toward the *perihelion*), the equation of the ellipse becomes

$$\frac{1}{r} = \frac{Z\,e^2m}{M^2}\,(1 + \epsilon\cos\omega). \tag{316}$$

The two constants M and ϵ are determined by the initial conditions; M represents the angular momentum[10] ($M = mr^2\dot{\omega}$), and ϵ the eccentricity (we are dealing with ellipses, for which $\epsilon < 1$). In order to find the semiaxes, we note that the maximum and minimum values of r, that is, the *aphelical* and *perihelical* distances (corresponding to $\omega = 180°$ and $\omega = 0°$), are

$$r_{\max} = \frac{M^2}{Z\,e^2m}\,\frac{1}{1-\epsilon}, \qquad r_{\min} = \frac{M^2}{Z\,e^2m}\,\frac{1}{1+\epsilon}, \tag{317}$$

and since the semimajor axis a is evidently given by $\frac{1}{2}(r_{\max} + r_{\min})$, we will have

$$a = \frac{M^2}{Z\,e^2m}\,\frac{1}{1-\epsilon^2}, \tag{318}$$

---

[10] We select the positive sense of ω coincident with the sense in which the ellipse is traversed, so that M is not negative.

while from geometry we have

$$b = a\sqrt{1 - \epsilon^2}. \tag{319}$$

The kinetic energy is

$$T = \frac{m}{2}(r^2 + r^2\dot{\omega}^2), \tag{320}$$

and the potential energy is $-Z\,e^2/r$. To calculate the total energy $E = T + U$ it is convenient (since this energy is a constant) to refer to a particular instant of the motion, selected in such a way as to simplify the calculation. For instance, for $r_{\min}$ we have $\dot{r} = 0$; then we find that

$$E = -\frac{1}{2}\frac{Z\,e^2}{a} \tag{321}$$

and note that the energy depends solely on the semimajor axis, and not on the semiminor axis.

(b) *Sommerfeld conditions.* We note that the system is doubly degenerate (since the three coordinates all vary with the same period). In order to select a system of coordinates (see §53) it is therefore necessary to take some perturbation into account: one of these is the relativistic correction applied to the laws of mechanics, and another might be the action of a weak external magnetic field. It is found that a system perturbed in this way is no longer degenerate and that the variables are separable if polar coordinates (in space) are taken with the pole at the nucleus and the polar axis directed along the field. Hence we shall adopt these coordinates even when neglecting the above-mentioned perturbations. Therefore let r be the radius vector, θ the colatitude (angle between the radius vector and the polar axis), and φ the longitude. The kinetic energy then has the known form

$$T = \frac{m}{2}(\dot{r}^2 + r^2\dot{\theta}^2 + r^2\sin^2\theta\dot{\varphi}^2) \tag{322}$$

and hence the momenta conjugate to r, θ, φ are, respectively,

$$\left.\begin{aligned} p_r &= m\dot{r}, \\ M_\theta &= mr^2\theta, \\ M_\varphi &= mr^2\sin^2\theta\dot{\varphi}. \end{aligned}\right\}$$

The last one has the mechanical meaning of "angular momentum with respect to the polar axis" or *projection upon the polar axis of the* *angular momentum* **M.** This projection is a constant in the absence of external forces.

We now may write the Sommerfeld conditions, which are

$$\oint p_r\,dr = m\oint \dot{r}\,dr = n'h, \tag{323}$$

$$\oint M_\theta\,d\theta = m\oint r^2\dot{\theta}\,d\theta = sh, \tag{324}$$

$$\oint M\,d\varphi = m\oint r^2\sin^2\theta\dot{\varphi}\,d\varphi = m_\varphi h, \tag{325}$$

where n', s, m_φ are three nonnegative integers.

Since M_φ is a constant, the last equation immediately gives

$$\pm 2\pi M_\varphi = m_\varphi h,$$

where the + or − signs hold according to whether the motion takes place in the direction of increasing or decreasing φ. Adopting the convention of indicating[11] by m^* the number $\pm m_\varphi$ (where the sign is chosen by the preceding criterion), we have

$$M_\varphi = m^*\frac{h}{2\pi}, \tag{325'}$$

which amounts to saying that the angular momentum with respect to the polar axis is expressed (in units $h/2\pi$) by an integer m^* ($\gtreqless 0$) which is called *magnetic quantum number* (and which corresponds to the magnetic quantum number m of the wave theory introduced in §46).

An analogous quantization for the (total) angular momentum $M = |\mathbf{M}|$ may be obtained from (324) and (325). We note first of all that upon equating the expressions (320) and (322) for the kinetic energy, and multiplying through by dt, we get the identity

$$r^2\dot{\theta}\,d\theta + r^2\sin^2\theta\dot{\varphi}\,d\varphi = r^2\dot{\omega}\,d\omega.$$

Then, adding the last two Sommerfeld conditions (324) and (325), and using the last identity, we obtain

$$m\oint r^2\dot{\omega}\,d\omega = (s + m_\varphi)h.$$

---

[11] For the present we shall adopt the notation m^* for the magnetic quantum number to avoid confusion with the electronic mass m. Later on, when there is no more reason for misunderstandings, we shall write m instead, as is customary.

We now recall that the angular momentum is given by

$$M = mr^2\dot{\omega}; \tag{326}$$

if we now define a new (nonnegative) quantum number k, putting

$$k = s + m_\varphi = s + |m^*|, \tag{327}$$

we may write[12]

$$\oint M\, d\omega = kh. \tag{328}$$

Since M is a constant, we obtain from (328)

$$M = k\frac{h}{2\pi}, \tag{329}$$

or: *the total angular momentum is an integral multiple of $h/2\pi$.* The quantum number k measuring this angular momentum in units of $h/2\pi$ was called *azimuthal quantum number* in the old Bohr-Sommerfeld theory. This designation is now reserved for the number l of wave mechanics, which, as we shall see later on, may be identified with $(k - 1)$.

(c) *Spatial quantization.* The last two Sommerfeld conditions determine the inclination of the orbital plane with respect to the polar axis, or else with respect to an external magnetic field (which may be very weak). In fact, given the angle α that the vector $\mathbf{M}$ (which is normal to the orbital plane) makes with this axis, we have

$$M_\varphi = M \cos \alpha;$$

and hence, substituting expressions (325′) and (329) for M_φ and M, we find (remembering, from what was said above, that $k \neq 0$)

$$\cos \alpha = \frac{m^*}{k}. \tag{330}$$

It follows that with k fixed, $\cos \alpha$ may take on only the series of discrete values $0, \pm 1/k, \pm 2/k, \ldots \pm 1$. However, in order to

---

[12] It is apparent that since M is the momentum conjugate to ω in the system of plane-polar coordinates (in the plane of the orbit), (328) may be interpreted as one of the two Sommerfeld conditions which would be obtained when the problem is treated in two dimensions, as if the electron moved in a given plane. The other is identical to (323). This way of treating the problem, that is, with the somewhat artificial restriction that the motion take place in a given plane, is the one usually followed in the elementary expositions of the Sommerfeld theory.

make our treatment agree with experimental facts, we must exclude the case $m^* = \pm k$, that is, $\cos \alpha = \pm 1$, in which case the plane of the orbit would be perpendicular to the magnetic field. This exclusion, for which the old Sommerfeld theory gives no satisfactory justification,[13] may be justified today by means of a comparison with the results of wave mechanics, of which this theory must represent an approximation (see §57 below). Assuming this for the moment, we arrive at the conclusion that the magnetic quantum number m^* takes on the $(2k - 1)$ values

$$-(k-1),\ -(k-2),\ \ldots\ -1,\ 0,\ 1,\ \ldots\ (k-1), \tag{331}$$

to which there correspond as many inclinations of the orbital plane.

The existence of these discrete inclinations is often denoted by the expression "spatial quantization."

It is apparent that the shape of the orbit does not enter in (324) and (325). Therefore, the results which we have deduced from these expressions (angular momentum quantization and spatial quantization) hold for any central motion, even if the force is not of the Coulomb type.

(d) *Shape of the orbits.* We must now take into account the remaining Sommerfeld condition (323), where n' ($= 0, 1, 2, \ldots$) is called *radial quantum number.*

The integral is calculated by taking ω as variable of integration and noting that

$$\dot{r} = \frac{dr}{d\omega}\dot{\omega} = \frac{M}{mr^2}\frac{dr}{d\omega},$$

so that we may write

$$\oint m\dot{r}\,dr = M\int_0^{2\pi}\frac{1}{r^2}\left(\frac{dr}{d\omega}\right)^2 d\omega = M\int_0^{2\pi} r^2\left(\frac{d\,1/r}{d\omega}\right)^2 d\omega;$$

substituting expression (316) for $1/r$, we obtain

$$M\epsilon^2\int_0^{2\pi}\frac{\sin^2\omega\,d\omega}{(1+\epsilon\cos\omega)^2}.$$

The integration may be effected by first integrating by parts and then making the transformation $\cos\omega = (1 - x^2)/(1 + x^2)$

---

[13] In the old quantum theory the value $m^* = 0$ was usually excluded, for reasons analogous to those which lead us to omit $k = 0$; the result, however, was not at all satisfactory (see Geiger and Scheel, *Hdb. der Physik* XIII, 1st ed., pages 145 and 164).

(or by a more rapid and elegant method due to Sommerfeld[14]), and we find

$$\oint m\dot{r}\,dr = 2\pi M\left(\frac{1}{\sqrt{1-\epsilon^2}} - 1\right)\cdot$$

Hence, taking into account (329), (323) yields the following condition for ϵ:

$$\frac{1}{\sqrt{1-\epsilon^2}} - 1 = \frac{n'}{k}$$

or else, by introducing a new integer (which is to be identified with the "principal quantum number" of §47),

$$n = k + n', \tag{332}$$

$$\sqrt{1-\epsilon^2} = \frac{k}{n}\cdot \tag{333}$$

Substituting this expression for $1 - \epsilon^2$ into (318), and (329) for M, we obtain for a the expression (depending only on n)

$$a_n = \frac{h^2}{4\pi^2 Z\, e^2 m}\, n^2. \tag{334}$$

Similarly, from (319), we see that the ratio of the semiaxes is

$$\frac{b}{a} = \frac{k}{n}\cdot \tag{335}$$

Generally the two integers k and n (rather than k and n') are used to characterize an orbit.

We see that k is less than or, at most, equal to n, since n' cannot be negative. We have $k = n$, that is, $n' = 0$, in the case of circular orbits (hence the n introduced into the Bohr theory, restricted to circular orbits, may be considered to be either an azimuthal or a principal quantum number). In addition, we must exclude the case $k = 0$. This again is a restriction which may be justified rigorously only by a comparison with the Schrödinger theory (see §57). In the old theory this restriction was justified by the argument that the case $k = 0$ would correspond to $b = 0$, as (355) shows; that is, the ellipse would degenerate into a line segment, and the electron would collide with the nucleus.

---

[14] See, for example, No. 18 of the Bibliography, page 655.

We then assume that the azimuthal quantum number k, with fixed n, may take on only the n values

$$k = 1, 2, \ . \ . \ . \ n, \tag{336}$$

and that to these there correspond as many ellipses, all having the same semimajor axis but different semiminor axes. The last one is the circle of radius a_n; of course all these ellipses have one focus at the nucleus (see Fig. 43).

(e) *Energy levels.* We note first that all ellipses corresponding to the same n and having the same a will have the same eneryg. Therefore the energy depends only on the principal quantum number n and not on k. Substituting (334) into (321), we obtain, for the energy of any one of the orbits of principal quantum number n,

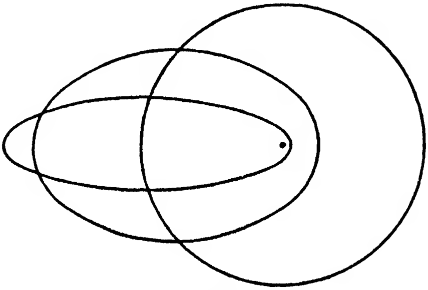

Fig. 43

$$E_n = -\frac{2\pi^2 Z^2\, e^4 m}{h^2}\,\frac{1}{n^2}, \tag{337}$$

an expression which is identical with that already found for the nth circular orbit in the Bohr theory (see §16 of Part I).

Hence the consideration of the elliptic orbits in hydrogen and in hydrogenlike systems does not add any new energy levels. For this reason the Bohr theory, although limited to circular orbits, already accounted for all observed spectral lines. The more complete theory, which has just been developed, tells us, however, that to each energy level there generally corresponds more than one type of motion because of the already mentioned degeneracy of the system. In fact, having fixed n (that is, the major axis), the azimuthal quantum number k may assume the values (336), and we therefore get different "shapes" of orbits. To each value of

k there correspond $(2k - 1)$ values of m^*, and hence each orbit may assume as many inclinations. Therefore the total number of orbits corresponding to a given principal quantum number n and hence to a given energy level (also called *statistical weight* or *multiplicity* of the level) is

$$\sum_{k=1}^{n} (2k - 1) = n^2.$$

Thus to the first level there corresponds one orbit, to the second level four orbits, and so forth. Every spectral line in general results from several kinds of quantum jumps. For instance, the first line of the Balmer series is emitted by all atoms in which the electron jumps from any one of the orbits $n = 3$, $k = 1, 2, 3$, to any one[15] of the orbits $n = 2$, $k = 1, 2$.

It is to be noted that a slight perturbation (for example, an electric or magnetic field) is sufficient to remove, totally or partially, the perfect overlapping of the energy values corresponding to the n^2 orbits of the same principal quantum number n. Each energy level will then split into a group of very closely spaced levels, corresponding to different values of k and of m^*.

57. Comparison with the Schrödinger theory. The theory of the hydrogen atom developed in the preceding section must be considered, as we know, as a first approximation of the Schrödinger theory, in the sense that the numerical values which are obtained from it for the various quantities having physical significance (energy, angular momentum, and so on) are approximately equal to those deducible from the rigorous theory developed in Chapter 8. However, the energy levels turn out to be exactly the same. The values of the angular momentum with respect to the axis, that is, of M_φ, are also exactly the same, provided that we assign to the magnetic quantum number m^* only the values (331) (excluding the values $\pm k$) and that we identify k with $(l + 1)$ of the wave theory (from which the exclusion of the value $k = 0$ follows). Hence the multiplicities of each level are also the same in the two theories.

On the other hand, the expression (329) of the angular momentum is only approximate. In fact, we have already pointed out

---

[15] Later it will be seen that some of these quantum transitions do not occur in reality, since they are excluded by selection rules.

that in the rigorous theory the latter is given (in units of $h/2\pi$), not by k but by $\sqrt{l(l+1)}$ or else by $\sqrt{k(k-1)}$, or $\sqrt{kl}$. In particular, for $k = 1$ we should get $M = 0$, whereas (329) gives $M = h/2\pi$.

Sometimes it may be convenient to substitute for (329) the formula

$$M = (k-1)\frac{h}{2\pi} = l\frac{h}{2\pi} \tag{329'}$$

that is, to measure M (in quantum units) by l rather than by k. Equation (329′) gives a slightly better approximation than (329) and immediately gives the right result for $k = 1$.

Let us now examine the physical significance which may be attributed to the orbits of the Bohr-Sommerfeld theory and to the motion of the electron along these orbits. In order physically to define the trajectory of the electron, as well as its motion, it would be necessary to be able to determine (at least conceptually) the position of the electron at successive instants, with uncertainties small compared with the dimensions of the orbit, and to repeat the observation many times upon the same atom. But in order to perform an observation of position with an uncertainty small compared with the radius a of the orbit,[16] it is necessary, by the uncertainty principle, to impart to the electron an indeterminate change in momentum of the order of magnitude $\Delta P \geq \sim h/a$. Hence in the succeeding observation the motion is already perturbed. We see, then, that the physical determination of the orbit is conceptually possible (approximately) only if the perturbation ΔP is negligible compared with the momentum P which the electron possesses in its orbital motion. Now, $P = \dfrac{M}{a} = \dfrac{h}{2\pi a}n$. Hence the condition is

$$\frac{h}{a} \ll \frac{h}{2\pi a}n \qquad \text{or} \qquad n \gg 2\pi.$$

Therefore we may say that the larger n is, the more exactly may an orbit be physically determined. We cannot attribute any physical significance to the first few orbits, for which $n = 1, 2, \ldots$.

In order to clarify what has been said above and to throw some light upon the relationship of the Bohr-Sommerfeld model to wave

---

[16] Here, for simplicity, we refer to circular orbits, but the reasoning may immediately be extended to elliptic orbits by replacing a and P by their average values.

mechanics, let us suppose that we wish to determine the successive positions of an electron within an atom, by irradiating it with radiation of wavelength short compared with the dimensions of the atom and observing the scattered light, just as in §23. (Obviously, we are dealing with idealized experiments.) If the atom is in one of the lower quantum states, its dimensions will be of the order of 10^{-8} cm, and therefore X rays will have to be used. But it is well known that one of the X-ray photons is sufficient to ionize the atom, and hence the electron is already expelled from the atom in the first observation, so that it is impossible to repeat the observation on the same atom. It is therefore not feasible to detect the orbit, but only *one* position of the electron with respect to the nucleus, and this operation may be performed with as great a precision as desired. If we then repeated this observation on many atoms (all in the same state), we could establish the statistics of the results obtained; or we could attribute to every point of space around the nucleus a "probability density" of finding the electron there. This is just the probability density which wave mechanics shows us how to calculate. Thus, for instance, if the atoms in question are all in their ground state ($n = 1$), the results of the position observations on the electron will be symmetrically distributed around the nucleus, becoming denser as we get closer to the nucleus, as shown by Fig. 39*a*.

If instead we were to consider atoms in a higher quantum state and hence of larger dimensions, we could be content to observe the position of the electron with less precision and hence with light of lower frequency, incapable of producing ionization. We can therefore think of repeating the observation many successive times on the same atom, thus approximately detecting the orbit of the electron and the law of its motion. But since, by the first observation, we have already perturbed the momentum (and the energy) of the electron in an uncontrollable way, we may not assign that orbit to a definite quantum state, and the indeterminacy of the state (that is, of n, k, m) will be the larger the more precisely the measurements of position have been carried out. This entire argument is faithfully represented in the scheme of wave mechanics. In fact, by superimposing a certain number of eigenfunctions corresponding to different values of n, k, m, we may build up as compact a wave packet as desired, which (see §26) moves around the nucleus and executes approximately the motion of the electron in the Bohr-Sommerfeld theory. This packet, if properly constructed, represents the probability density at every instant and hence permits the results of successive position observations to be predicted statistically. If these observations are performed with great precision (that is, if we make our observations with radiation of very short wavelength, compatible with the necessity of avoiding ionization), we must construct a very small wave packet, for which it will be necessary to superimpose eigenfunctions belonging to a wide range of values for n, k, m; that is, we must leave the quantum state highly indeterminate. Hence we see that the concept of "orbit" and that of "quantum state" are *complementary* in the sense of Bohr (see §22), just as are those of position and velocity, and so on.

Evidently, then, it is not possible to construct a sufficiently small wave packet if we want to use only the first three or four eigenfunctions, for example. This impossibility corresponds to the fact, already mentioned, that the orbits corresponding to the lowest values of n have no physical significance.

58. Motion of the nucleus. An important refinement which must be made in the preceding theory of hydrogenlike systems consists in accounting for the fact that the nucleus is not strictly fixed, as we have assumed up to now, but describes a small orbit about the center of gravity of the system. As is known from mechanics, the two-body problem may be reduced to that of a single body attracted by a fixed center, provided that we modify the mass of the moving body slightly. The motion of an electron of mass m with respect to the nucleus of mass M is governed by the same equations as the motion, with respect to a fixed nucleus, of a particle of mass (*reduced mass*)

$$m' = \frac{m}{1 + \frac{m}{M}}. \tag{338}$$

The Sommerfeld conditions are also obtained from those for the fixed nucleus by simply substituting m' for m. This fact is understood most easily if we take for generalized coordinates of the system the polar coordinates of the nucleus with respect to the center of mass (r_0, θ_0, φ_0), and the polar coordinates of the electron with respect to the nucleus (r, θ, φ). We then find that the momenta conjugate to the former are identically zero, and those conjugate to the latter three are

$$p_r = m'\dot{r}, \qquad M_\theta = m'r^2\dot{\theta}, \qquad M_\varphi = m'r^2 \sin^2\theta\dot{\varphi};$$

Hence, of the six Sommerfeld conditions, the first three are identically satisfied, and the other three coincide with (323), (324), (325) except for the substitution of m' for m.[17] The energy levels are therefore given by

$$E_n = -\frac{2\pi^2 Z^2 e^4 m'}{h^2}\frac{1}{n^2} = -Z^2\frac{R_M hc}{n^2},$$

where we have put (in analogy to the expression for the Rydberg constant given in §16 of Part I):

$$R_M = \frac{2\pi^2 e^4 m'}{ch^3} = \frac{R}{1 + m/M}, \tag{339}$$

and the spectral terms are

$$\tau_n = Z^2\frac{R_M}{n^2}.$$

Therefore the effect of the motion of the nucleus amounts to replacing the Rydberg constant R by the slightly smaller constant R_M. This correction, which depends on the mass M, is slightly different for different hydrogenlike systems. For instance, we find that

for H:	$R_{\mathrm{H}} = 109{,}677.58\ \mathrm{cm}^{-1}$,
for He$^+$:	$R_{\mathrm{He}^+} = 109{,}722.26\ \mathrm{cm}^{-1}$.

These results indicate that the frequencies of the same line in the various hydrogenlike spectra should not be exactly proportional to Z^2. For example, the even lines of the Pickering series will not exactly coincide with those of the Balmer series (as would be expected from the first approximation of §16 of Part I) but will be slightly displaced toward the violet. This is actually the case, and the difference (which is of the order of one angstrom) is perfectly observable in the spectrum given by a mixture of hydrogen and helium. In fact, the measurement of these differences constitutes

---

[17] Note that M still has the significance of total angular momentum of the system; hence the azimuthal quantum number k retains its meaning.

one of the most exact methods for the determination of the ratio of the mass of the electron to the mass of the nucleus. The value of R given by (21) of §16 is the one that would correspond to a nucleus of infinite mass, and is therefore sometimes designated by R_∞.

An entirely analogous correction may be made in the Schrödinger theory and will lead to the same result (see §21 of Part III).

59. Atoms with several electrons.[18] For atoms (or ions) with more than one electron, the Bohr-Sommerfeld method would lead first of all to a consideration of the mechanical problem of the motion of the electrons under the action of nuclear attraction and their mutual repulsions. (This repulsion is not quite negligible, as is the attraction of the planets in the astronomical problem, to a first approximation.) The solution of this problem is in general practically impossible. But even where an approximate solution has been found (in the case of the helium atom, with two electrons), it was realized that by quantizing its motion by the Sommerfeld conditions, energy levels which were grossly in error were obtained. Such a result implies that in this case the approximation given by the first two terms of (295) is entirely inadequate. Nevertheless, a series of qualitative results of the greatest importance was obtained by means of the following considerations.

Since the spectral terms of many atoms and ions may be represented in the Rydberg or the Ritz forms (see §15 of Part I) which are analogous to the Balmer form, Bohr was led to consider that the emission of these spectra is produced by the quantum jump of only one of the electrons, while the other electrons retain their orbits almost unchanged. Thus we see the possibility of considering (as an artifice) the atom or the ion to be composed of two parts. One part, consisting of the nucleus and all the electrons but one, is called *core* or *remainder of the atom*, and is considered to have an invariant structure, to a first approximation. The other part, consisting of the remaining electron, which is called *emission electron*, may travel on various quantum orbits under the action of the forces exerted by the core. This electron is usually one of the so-called *valence electrons*, that is, one of the electrons which,

---

[18] See a treatise on spectroscopy (e.g., G. Herzberg: *Atomic Spectra and Atomic Structure*, New York: Dover, 1944), or else No. 23, 27, or 32b of the Bibliography.

because they are less strongly bound to the atom, may be more easily removed from the atom, when it is transformed into a positive ion.

The forces which the core exerts upon the emission electron, though naturally rather complex, may be schematized by considering all the electrons of the core as forming a continuous distribution or "cloud" of negative electricity around the nucleus, which is static and spherically symmetrical so as to give rise to a central force field which is superimposed upon the field due to the nucleus.[19] We know that such a field, for points outside of the cloud, is the same as if it were concentrated at the center, or else as if the positive charge of the nucleus were diminished by an amount equal to the agglomerated charge of the electrons constituting the cloud. If there are α electrons, the charge of the nucleus is effectively reduced from Ze to $(Z - \alpha)e$ (*effective charge*). Therefore the emission electron is attracted toward the core (as long as it is outside the core) by a force $(Z - \alpha)\, e^2/r^2$ instead of $Z\, e^2/r^2$. Hence it is said, somewhat inaccurately, that the electrons of the core exert a *screening action* upon the nuclear attraction, and the number α is called the *screening constant.* If, for instance, we are dealing with a neutral atom, we shall have $\alpha = (Z - 1)$, and hence the effective charge will be e. If we are dealing with an atom ionized z times, α will be $\alpha = (Z - 1 - z)$, and hence the effective charge will be $(z + 1)e$.

If, however, the emission electron penetrates into the core, the screening action will decrease, because, as we know, the field produced by a spherical distribution of electricity at a point within it, at a distance r from the center, is the same as would be produced by that portion of the electric charge which is contained within the sphere of radius r, whereas the external charge remains ineffective. Hence for an electron penetrating the core, the effective charge of the nucleus gradually increases with decreasing r and approaches the true charge Ze when r approaches zero. In that case, then, we may say that the emission electron moves in a field which is Newtonian outside of a certain sphere (boundary of the core) and

---

[19] According to wave mechanics, this idealization now appears less arbitrary than one might think. In fact, very often the Schrödinger ψ-function corresponding to the electrons of the core is such that $|\psi|^2$, or *mean* electric charge density, is independent of time and is spherically symmetrical.

non-Newtonian inside. The law of variation of the field within the core depends on the law with which the electric charge density is assumed to be distributed there. It may be idealized in various ways, according to the degree of approximation desired.

The calculation of the energy levels of an atom with several electrons therefore reduces, in this approximation, to the calculation of the energy levels of a single electron in a non-Newtonian field. This problem will be attacked in a manner analogous to that of hydrogenlike systems. The ordinary mechanics of central motions will allow us to determine a continuous infinity of motions. It turns out that these motions may, to a first approximation, be considered as Keplerian motions upon which there is superimposed a slow, uniform rotation, called *precession,* so that the trajectory has the shape of Fig. 44 and is called a *rosette.* Then imposing the three Sommerfeld conditions, we find the quantum orbits, which in this case also happen to depend upon three integers: an *azimuthal quantum number* k ($= 1, 2, \ldots$) or l ($= 0, 1, 2, \ldots$) which measures the angular momentum in units of $h/2\pi$, a *magnetic quantum number* m ($= 0, \pm 1, \ldots \pm l$) expressing the projection of this angular momentum upon an axis, and a *principal quantum number* n. In general, however, circumstances will not be such

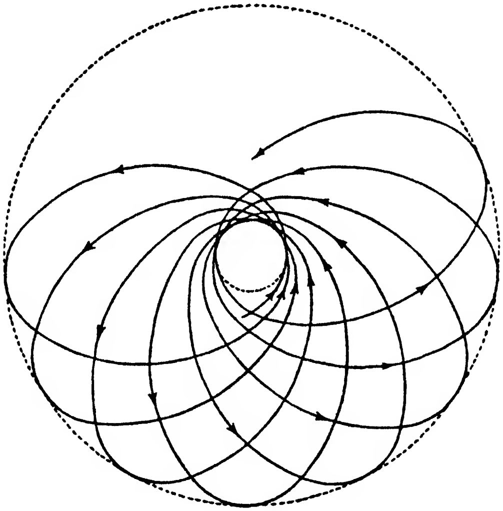

Fig. 44

that the energy will depend only on n, as in the hydrogenlike systems, but rather will depend on n and l (not on m, by reason of the spherical symmetry of the field[20]), and therefore the energy levels as well as the spectroscopic terms will constitute a series with *double index* rather than with a single index. That is, we shall write E_{nl}, τ_{nl}. This is the most important result of these considerations and retains its value even in wave mechanics. It may be added that, by judiciously specifying the field created by the core, Sommerfeld has succeeded in finding the Rydberg form for spectral terms, and in a further approximation, the Ritz form.[21] It is now evident that if the orbit of the emission electron were lying entirely outside the core, where the field is Newtonian (*nonpenetrating orbit*), we should again find the energy levels of hydrogen and hence the Balmer terms. In this way there is justification of the experimental fact that the more the terms of any atom or ion in general approach the Balmer form, the higher are their quantum numbers.

In spectroscopy it has become conventional to indicate the spectral terms corresponding to the energy levels of an electron by means of a small letter designating the value of the azimuthal quantum number l, preceded by a number indicating the principal quantum number, rather than by the notation τ_{nl}. The azimuthal quantum number is designated by means of the following letters, derived from the terminology of the old spectroscopy (we shall also give the value of k, the azimuthal quantum number in the Sommerfeld theory):

$l =$	0	1	2	3	4	5 . . .
$k =$	1	2	3	4	5	6 . . .
Letter used:	*s*	*p*	*d*	*f*	*g*	*h* . . .

(Terms with higher l rarely occur.) Hence, for example, the term for which $n = 3$ and $l = 0$ is indicated by 3*s* rather than by τ_{30}, and we speak of *terms of the s-series, of the p-series*, and so on, or also of *s-terms, p-terms*, and so on. It is apparent that since always $n \geq l + 1$, the *s*-series starts with the term 1*s*, the *p*-series with the term 2*p*, the *d*-series with the term 3*d*, and so on.

---

[20] The energy will depend on m when the atom finds itself in a magnetic field of an intensity sufficient to perturb the motion. The Zeeman effect is then produced.

[21] See No. 18 of the Bibliography.

Graphically, the levels corresponding to the terms are usually represented in different columns: one for the s-series, one for the p-series, and so on, as shown on the right of Fig. 45.

In hydrogenlike systems, the levels of the various columns would all coincide (in our approximate treatment) and therefore are represented in a single column, as in Fig. 8 or on the left of Fig. 45.

In the atoms and ions which are not hydrogenlike, the atoms which jump from a state of principal quantum number n to one of principal quantum number n' do not all emit the same spectral line, but different lines according to the initial and final values of the azimuthal quantum number k. In the case of hydrogenlike systems these lines coincide, and the particular simplicity of these spectra results from this fact. However, the coincidence is not perfect, as we shall see later on.

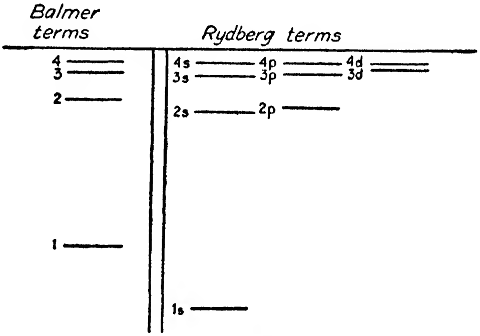

Fig. 45

60. Note on the relativistic theory of hydrogenlike systems. The motion of the emission electron takes place with a velocity of the order of 10^8 cm/sec, that is, of the order of a few thousandths of the velocity of light, as may be easily calculated. Hence ordinary mechanics may be applied to the electron only approximately, and it is to be expected that a more rigorous treatment, using relativistic dynamics, will apply corrections to the previously found results which are not entirely negligible.

If we repeat the calculations of §56 for hydrogenlike systems but make use of relativistic mechanics instead of ordinary mechanics,[22] we find first of all that the electron describes an orbit in the

---

[22] See, for instance, No. 18 of the Bibliography.

shape of a rosette; that is, it describes an ellipse with one focus at the nucleus, according to Kepler's laws. This ellipse, however, rather than being stationary, rotates slowly about the nucleus. The effect of the relativistic correction on the orbit is therefore analogous to the effect which, for nonhydrogenlike atoms, is due to the fact that the field of force is not exactly Newtonian. Hence the orbit is no longer a closed curve, nor the motion periodic. Consequently, the relativistic perturbation partially removes the degeneracy of the system, as has been pointed out in §56. As far as the energy levels are concerned, we find the following approximate expression:

$$E_{nk} = -\frac{2\pi^2 Z^2 e^4 m}{h^2}\frac{1}{n^2} - \frac{\pi^4 Z^4 e^8 m}{c^2 h^4}\frac{1}{n^3}\left(\frac{8}{k} - \frac{6}{n}\right);$$

or else, if we recall the expression for the Rydberg constant, found in §16 of Part I, and introduce the constant[23] (called the *fine-structure constant*),

$$\alpha = \frac{2\pi e^2}{hc} = 7.2969 \times 10^{-3} = \frac{1}{137.04} \tag{340}$$

and the last expression becomes

$$E_{nk} = -Z^2\frac{Rhc}{n^2}\left[1 - \frac{Z^2\alpha^2}{n^2}\left(\frac{3}{4} - \frac{n}{k}\right)\right]. \tag{341}$$

Neglecting the term in α^2, we have here the expression already found in the nonrelativistic theory. This term represents a correction whose ratio to the principal term is of the order of α^2, or a few hundred-thousandths. This correction, as we see, is dependent not only on n but also on k. Hence all the orbits having the same principal quantum number n and azimuthal quantum numbers $k = 1, 2, \ldots n$ (that is, the ellipses with the same major axis and different eccentricities) do not exactly correspond to the same energy level, as was found in a first approximation, but to different levels, which lie very close together. Therefore we speak of *multiplet levels.* Of course to this multiplicity of levels there corre-

---

[23] This quantity, which occurs also in connection with the spinning electron, has the dimensions of a pure number and is equal, as may be easily verified, to the ratio of the electron velocity in the first circular orbit to the velocity of light. The numerical value given here is quoted by R. T. Birge, *Phys. Rev.* **79**, 193 (1950).

sponds a multiplicity of spectral lines. The atoms which jump from a state of principal quantum number n, with various values of k, to a state of principal quantum number n', with different values of k', emit lines which do not coincide exactly. Therefore, strictly speaking, the same effect occurs in hydrogenlike systems as has been discussed in the previous section for more complex atoms, although for a different reason. The frequency differences in question are so small, however, that they may be detected only with instruments of high resolving power and only if the lines are very narrow. Therefore the group of very closely spaced lines will be considered as a "line" composed of several "components" (*fine-structure*).

Formula (341) permits us to calculate the frequencies of the various components of each line. Here we shall restrict ourselves to noting that the frequency differences are of the order of $\alpha^2 Z^4 R$, and therefore must be 16 times as large in the spectrum of ionized helium ($Z = 2$) as in the corresponding hydrogen lines. This fact, together with the circumstance that it is easier to obtain narrow lines with helium than with hydrogen, is the reason why the fine-structure is easier to observe in helium.

The predictions which are made by formula (341) (supplemented by a *selection rule* which will be stated in §64) agree almost exactly with the results of observations carried out with helium and with hydrogen. Nevertheless, this coincidence must be considered fortuitous to a certain extent.

In fact, the preceding theory neglects two classes of facts. First, the Bohr-Sommerfeld model represents only a first approximation; secondly, it neglects the influence of the electron spin. In Chapter 14 it will be seen that when these two errors are eliminated, by using the relativistic wave mechanics of Dirac (which implicitly accounts for spin), the result obtained is fairly close to the one in formula (341) (and exactly confirmed by experiment), so that the partial success of (341) is due to the fact that the errors depending upon the two above-mentioned causes almost exactly cancel each other.

61. The Bohr magneton. An important consequence of the Rutherford model of the atom is that an atom must in general possess a magnetic moment because of the orbital motion of the electrons, whose trajectories are equivalent to as many electric circuits. On the basis of this model, let us calculate the magnetic moment produced by the motion of the electron in the case of hydrogenlike systems.

To make this calculation we note that if τ is designated as the period of the Keplerian motion, the charge e passes $1/\tau$ times per second through any point of the trajectory. This motion is as if the ellipse were traversed by a current of intensity e/τ. This ellipse, according to a well-known principle of electromagnetic theory, is equivalent to a magnetic shell whose potential (magnetic moment per unit area) is $(1/c)(e/\tau)$. Hence, if the area of the ellipse is called S, its magnetic moment is

$$\mu = \frac{Se}{c\tau}. \tag{342}$$

Now, by the law of areas,

$$dS = \frac{1}{2} r^2 \, d\omega = \frac{M}{2m} \, dt,$$

and hence, by integration over a whole period,

$$S = \frac{M}{2m} \tau. \tag{343}$$

Substituting into (342), we see that the magnetic moment μ turns out to be related in magnitude to the angular momentum M by

$$\mu = \frac{e}{2mc} M. \tag{344}$$

If we then consider that the vectors $\boldsymbol{\mu}$ and $\mathbf{M}$ are perpendicular to the orbital plane and are directed in opposite directions (as may be readily seen), we may also write the vector relation

$$\boldsymbol{\mu} = -\frac{e}{2mc} \mathbf{M}, \tag{344'}$$

where e stands for the electronic charge in absolute value. This result may be extended to systems with as many electrons as desired.

We now recall that M is always an integral multiple of $h/2\pi$, according to (329) or preferably (329′), so that we may write

$$\mu = l \frac{eh}{4\pi mc},$$

and we find that the magnetic moment due to the orbital motion of electrons is always an integral multiple of

$$\mu_0 = \frac{eh}{4\pi mc}. \tag{345}$$

This elementary magnetic moment is called the *Bohr magneton.* Its value, in electromagnetic c.g.s. units, is

$$\mu_0 = 0.92736 \times 10^{-20}.$$

Moreover, it is apparent that, because of the result found in §56' (spatial quantization) which may be immediately extended to nonhydrogenlike systems, an atom placed in a magnetic field orients itself in such a way that the component of **M** along the direction of the magnetic field is $m\,(h/2\pi)$, where m is an integer which we have called *magnetic quantum number.* From (344′) we may then see that the component of the magnetic moment in the direction of the field will also be quantized; that is,

$$\mu_H = m\mu_0. \tag{346}$$

Stern and Gerlach have given a remarkable experimental demonstration of this fact, which has made it possible to measure the magnetic moments of different atoms directly.[24]

Incidentally, it is to be noted that the result (346) may also be obtained by wave mechanics, as has been shown by Fermi. In fact, we may use expression (137′) from §31 for the average electric current density j and substitute for u the expression found in §46 for the electron under the action of a central field of force, with the result

$$\psi = \frac{1}{\sqrt{2\pi}} R(r)\Theta(\theta)\, e^{im^*\varphi},$$

where R and Θ are two functions whose exact form is not of interest; it suffices to mention that they are real and normalized according to formulas (244) and (252). We readily find that the components of **j** along the radius vector and along the meridian are zero, whereas the component along increasing φ is

$$j_\varphi = \frac{eh}{4\pi im}\left(\psi^* \frac{\partial\psi}{\partial\varphi} - \psi\frac{\partial\psi^*}{\partial\varphi}\right)\frac{1}{r\sin\theta} = \frac{eh}{4\pi^2 m}\,\frac{R^2\Theta^2}{r\sin\theta}\, m^*.$$

---

[24] See, for example No. 27 of the Bibliography, page 168.

The surface element $r\,dr\,d\theta$ of the meridian plane is therefore traversed by a current of intensity $j_\varphi r\,dr\,d\theta$, which describes a circle of radius $r \sin\theta$ and is thus equivalent to a magnetic shell of magnetic moment

$$d\mu_H = \frac{\pi}{c} r^2 \sin^2\theta\, j_\varphi r\,dr\,d\theta.$$

Integrating over the entire half-plane of the meridian, we obtain the total magnetic moment in the direction of the polar axis, which is

$$\mu_H = \frac{eh}{4\pi mc} m^* \int_0^\infty R^2 r^2\,dr \int_0^\pi \Theta^2 \sin\theta\,d\theta;$$

and since, by virtue of the quoted normalization conditions, the two integrals are equal to 1, again we find, recalling (345), the result of (346).

62. The spinning electron and the inner quantum number. The considerations of the preceding section led for a long time to the belief that all phenomena depending upon the magnetic moment of the atoms (such as paramagnetism and ferromagnetism) could be explained by the moment due to the orbital motions of the electrons. But there are many phenomena (Zeeman effect, gyromagnetic effects) which permit the determination of the ratio μ/M of the magnetic moment to the mechanical moment (that is, the angular momentum) of the atom. According to the preceding section, this ratio should prove to be constant and equal to $e/2mc$. Instead it was found that this gyromagnetic ratio sometimes has different values, but is always a simple multiple of $e/2mc$. The most obvious explanation is that some part of the atom has a mechanical and a magnetic moment which are of an origin different from that of the orbital motion, and that the ratio of these moments is different from the one mentioned above. It has been found that these difficulties disappear if it is assumed that each electron possesses an intrinsic angular momentum (*spin*) equal to a *half* quantum unit, namely,

$$M_0 = \frac{1}{2}\frac{h}{2\pi}$$

(as if it rotates about itself like a top), and an intrinsic *magnetic moment* pointing in the opposite direction and having the value of one Bohr magneton, namely,

$$\mu_0 = \frac{eh}{4\pi mc}.$$

This is the hypothesis of the *spinning electron,* which we have mentioned in §25 of Part I, where the reason for this improper designation has been explained. It is apparent that the ratio μ_0/M_0 is twice the corresponding ratio relative to orbital motion.

Of course the orientation of the axis of the spinning electron in a magnetic field cannot be deduced from the preceding hypothesis. It requires a new hypothesis, which is suggested by the result of §56 on the spatial quantization of the orbits, and represents the natural generalization thereof. In order to state the spatial quantization rule in its most general form, let us consider a system in which the total angular momentum is represented by a vector $\mathbf{j}$. Using quantum units, as we shall always do in this section, j will be an integer or a half-integer (that is, an integer plus $\frac{1}{2}$). *We shall assume that this system, when placed in a magnetic field, will orient itself in such a way that the projection of* $\mathbf{j}$ *along the direction of the field will have one of the* $2j + 1$ *discrete values lying between* $-j$ *and* $+j$ *(extremes included) and spaced apart by unity;* that is,

$$-j,\ -(j-1),\ .\ .\ .\ (j-1),\ j. \tag{347}$$

The extreme values evidently correspond, in the intuitive model, to $\mathbf{j}$ antiparallel or parallel to the field. Naturally, if j is an integer, the values of the series (347) will be integers. If j is a half-integer, they will be half-integers.

This rule, when applied to the orbital angular momentum of the electron, brings us back to the result of §56. If it is applied instead to the angular momentum due to its spin alone, in which case $j = \frac{1}{2}$, it evidently yields the following result: the projection of the spin in the direction of the field can have only the two values $\pm\frac{1}{2}$. In other words, *the electron spin is always either parallel or antiparallel to the field.*

Chapter 15 explains how the values for M_0 and μ_0 and their projection on the field direction may be deduced from quantum mechanics without the introduction of new hypotheses. For the present, however, we shall assume these values as hypotheses justified a posteriori by experience.[25]

It is now apparent that the electron of hydrogenlike systems, because of its orbital motion, finds itself in a magnetic field perpendicular to the plane of the orbit. In order to understand this concept, we should think of an observer who travels with the electron; he would see the nucleus revolving about the electron and hence (since an electric charge in motion is equivalent to a current) would see himself encircled by an electric current, and therefore located in a magnetic field perpendicular to the plane of that current.[26] Because of the effect of this field, the spin will always orient itself normally to the orbital plane, in one of two directions, and hence the intrinsic angular momentum of the electron will be added to (or subtracted from) the angular momentum of the orbital motion, which is equal to l in our approximation. The resultant angular momentum will therefore be

$$\begin{aligned} j &= l \pm \tfrac{1}{2} \qquad && \text{if } l \neq 0 \ (p, d, f, \text{ orbits}), \\ j &= \tfrac{1}{2} && \text{if } l = 0 \ (s \text{ orbit}). \end{aligned}$$

In order to distinguish the two states of the atom corresponding to the two possible spin orientations, it suffices to add to the indication of the three quantum numbers n, l, m, which characterize the orbit, the indication of the resultant angular momentum, which is denoted by j, expressed in quantum units, and is called *inner quantum number.* In the case we are considering, this number may take on only the two (half-integral) values $j = l \pm \frac{1}{2}$ if $l \neq 0$, and $j = \frac{1}{2}$ if $l = 0$. In the latter case, furthermore, we are not to consider the two spin orientations as corresponding to two distinct states of the atom, because, since the vector **M** (and consequently also the magnetic field) is zero, the two orientations which the spin may assume are not physically distinct.

---

[25] It will further be seen that the "vector model" of which we shall make use (vector susceptible of discrete orientations) does not represent the properties of spin at all well. Nevertheless, like all models, it is very useful in aiding intuition and in simplifying the terminology.

[26] We can reach the same conclusion from the theory of relativity, in which it will be remembered that if an electric field exists in a certain frame of reference that is assumed to be fixed, then an electric and a magnetic field will exist in another system that is in motion with respect to the first. In our case, the electric field of the fixed system is the one produced by the nucleus.

All that has been said applies not only to hydrogenlike systems but also to the emission electron of the alkali metals, because in these atoms the magnetic fields produced by the $Z - 1$ electrons of the atomic core cancel each other. However, in other atoms the spins and magnetic fields produced by the electrons (or by some of the electrons) of the core must also be considered, and hence the possibilities are more numerous. For these arguments we refer to a volume on spectroscopy (see footnote 18 on page 252). We confine ourselves here to stating that the compound angular momentum of all the electrons of the atom (due either to their orbital motion or to their spin) is a constant vector whose magnitude is generally expressed (in quantum units) by an integer or half-integer J, which is called *inner quantum number* of the atom. This number is identical with the inner quantum number j of the emission electron in the case of the alkali metals and hydrogenlike systems.

Now turning to hydrogenlike systems and to alkali metals, we note that to the two values of the inner quantum number j there correspond two slightly different energy levels. In fact, by likening the electron to a magnet of moment $\boldsymbol{\mu}_0$, we recognize that the latter, finding itself in a field $\mathbf{H}$ due to the effect of its orbital motion, possesses a potential energy $-\boldsymbol{\mu}_0 \cdot \mathbf{H}$. Hence to the two orientations which the spin may assume there correspond the two values

$$\epsilon_1 = \mu_0 \bar{H}, \qquad \epsilon_2 = -\mu_0 \bar{H},$$

of the magnetic energy, which is to be added to the kinetic energy and the potential energy of the atom. We have indicated by $\bar{H}$ the average value of the field H along the orbit. However, a more thorough examination, taking into account the theory of relativity, shows that the preceding values must be cut in half.[27] By means of a calculation not given here, we find, making use of the known expressions for the Rydberg constant R and the fine-structure constant α of formula (340),

$$\epsilon_1 = \frac{1}{2}\alpha^2 Z^4 \frac{Rhc}{n^3 l^2}, \qquad \epsilon_2 = -\frac{1}{2}\alpha^2 Z^4 \frac{Rhc}{n^3 l^2}. \tag{348}$$

Every energy level is therefore split into two neighboring levels by the effect of the spin. Their difference is of the order of $\alpha^2 Z^4 Rhc$,

---

[27] See L. H. Thomas, *Phil. Mag.* **3**, 1 (1927); J. Frenkel, *Zeits. f. Physik* **37**, 243 (1926).

like the one produced by the relativistic correction [see (341)]. We shall not insist upon the quantitative aspects of this theory, however, since quantum mechanics appreciably modifies formulas (348), as we shall see.

63. The correspondence principle. We shall now make a comparison between the spectrum which a system emits according to the Bohr-Sommerfeld theory (*quantum-mechanical spectrum*) and the spectrum (which we shall call *classical*) which it would emit if, though able to exist only in quantum states, it were to radiate, in each of these states, in accordance with the ordinary laws of electromagnetic theory. First, then, we shall see what would be the emitted frequencies in the latter case (which is, of course, entirely fictitious).

We recall that according to the laws of electromagnetic theory, the emission of radiation by a system of electric charges is determined by the variation of its "electric moment," which is a function of the generalized coordinates q, which in turn are periodic functions of time, each with a frequency $\nu_i = 1/T_i$. It follows that each of the components X, Y, Z of the electric moment may be developed in a *multiple Fourier series*, of the form

$$X(t) = \sum_{\substack{\tau_1, \tau_2 \ldots \tau_f \\ -\infty}}^{\infty} A_{\tau_1, \tau_2 \cdots \tau_f} e^{2\pi i(\tau_1 \nu_1 + \cdots + \tau_f \nu_f)t}, \tag{349}$$

and similarly for the Y and Z components.

In fact, let us for an instant consider X to be a function of the coordinate q_1 alone, and let us keep the other coordinates constant. X will then be a periodic function of time, with frequency ν_1, and may be developed in a simple Fourier series,

$$X = \sum_{\tau_1 = -\infty}^{\infty} a_{\tau_1} e^{2\pi i \tau_1 \nu_1 t},$$

where the coefficients a_{τ_1} are functions of $q_2, q_3, \ldots q_f$. We may now apply the same procedure to each of these coefficients, considering them to be functions of q_2 only and leaving $q_3, \ldots q_f$ constant. We then will have

$$a_{\tau_1} = \sum_{\tau_2 = -\infty}^{\infty} b_{\tau_1 \tau_2} e^{2\pi i \tau_2 \nu_2 t};$$

and hence, substituting into the expression for X, the latter becomes the double series

$$X = \sum_{\substack{\tau_1\tau_2 \\ -\infty}}^{\infty} b_{\tau_1\tau_2} e^{2\pi i(\tau_1\nu_1+\tau_2\nu_2)t},$$

where the coefficients $b_{\tau_1\tau_2}$ depend on $q_3, \ldots q_f$.

Applying the same process successively, we evidently obtain the expansion (349).

Now if the components of the electric moment can be decomposed into a sum of sinusoidal terms, each component of the electric field and the magnetic field which it generates may also be decomposed in the same way. Hence the radiation emitted by the system in the state in question consists of the superposition of monochromatic radiations, whose frequencies (which we shall indicate by ν_{cl}) are given by all the values taken on by the expression

$$\nu_{\text{cl}} = \tau_1\nu_1 + \tau_2\nu_2 + \cdots + \tau_f\nu_f, \tag{350}$$

where we assign to $\tau_1, \tau_2, \ldots \tau_f$ all integral values (positive or negative) which do not make the right-hand side negative. The intensity of each of these monochromatic components is proportional to

$$(A_{\tau_1\tau_2\ldots\tau_f})^2 + (B_{\tau_1\tau_2\ldots\tau_f})^2 + (C_{\tau_1\tau_2\ldots\tau_f})^2, \tag{351}$$

where we have designated by $B_{\tau_1\tau_2,\ldots\tau_f}$ and $C_{\tau_1\tau_2,\ldots\tau_f}$ the coefficients of the expansion of Y and Z analogous to (349). In particular, if for some of the groups of values $\tau_1\tau_2, \ldots \tau_f$ all three coefficients A, B, C vanish, the corresponding monochromatic component will be missing from the emitted radiation.

The fictitious spectrum which we have agreed to call classical is therefore composed of lines which are specified by two groups of indices: $n_1, n_2, \ldots n_f$ (which define the state of the system, and hence the frequencies $\nu_1, \nu_2, \ldots \nu_f$) and $\tau_1, \tau_2, \ldots \tau_f$. Therefore, to each line of the classical spectrum we may *make correspond* a line of the quantum spectrum, namely, that line which is emitted in the transition from the state with quantum numbers $n_1, n_2, \ldots n_f$ to the state with quantum numbers $n_1 - \tau_1, n_2 - \tau_2, \ldots n_f - \tau_f$. We shall now ascertain the frequency of this line when it is calculated according to the quantum theory.

To do this, we first recall from rational mechanics (see §52) that the energy E of a multiply periodic system may be expressed in the form

$$E(J_1, J_2, \ldots J_f), \tag{352}$$

that is, as a function of the phase integrals J_i (which replace the f constants of integration α_i). It may then be shown that the partial derivatives of this function are equal to the frequencies ν_i corresponding to the several degrees of freedom:

$$\frac{\partial E}{\partial J_i} = \nu_i \qquad (i = 1, 2, \ldots f). \tag{353}$$

With this assumption, in the quantum jump mentioned above the phase integrals will go from the values $J_i = n_i h$ to the values $(n_i - \tau_i)h$. Hence the energy emitted is

$$\Delta E = E(J_1, J_2, \ldots J_f) - E(J_1 - \tau_1 h, J_2 - \tau_2 h, \ldots J_f - \tau_f h);$$

or, when the mean-value theorem is applied,

$$\Delta E = \overline{\frac{\partial E}{\partial J_1}} \tau_1 h + \overline{\frac{\partial E}{\partial J_2}} \tau_2 h + \cdots + \overline{\frac{\partial E}{\partial J_f}} \tau_f h,$$

where the bars are to indica e that the derivatives refer to a state (not quantized) intermediate between the initial and final states. Keeping (353) in mind, we may write, with similar meaning for the bars,

$$\Delta E = \bar{\nu}_1 \tau_1 h + \bar{\nu}_2 \tau_2 h + \cdots + \bar{\nu}_f \tau_f h.$$

Hence the frequency emitted in the quantum jump in question is

$$\nu_{\text{qu}} = \frac{\Delta E}{h} = \bar{\nu}_1 \tau_1 + \bar{\nu}_2 \tau_2 + \cdots + \bar{\nu}_f \tau_f. \tag{354}$$

A comparison of this formula with (350) shows that if the initial and final states differ slightly (that is, if the numbers τ_i are small compared with the n_i, implying that the n_i must be large), the frequencies $\bar{\nu}_i$ very closely approach the frequencies ν_i corresponding to the initial state, and hence the frequency ν_{qu} of the line of the quantum spectrum becomes approximately equal to the frequency ν_{cl} of the corresponding line in the classical spectrum.

Therefore we should say that *in the limit of high quantum numbers, the "corresponding" lines* (in the sense explained above) *have* *the same frequency.*[28] This theorem, which we have derived from the fundamental postulates of the Bohr-Sommerfeld theory, proves that this theory satisfies (as far as the prediction of spectral frequencies is concerned) the essential condition required of any atomic mechanics, namely, to have ordinary mechanics and electromagnetic theory for a limit.

All that has now been said refers only to the frequencies of spectral lines, not to their intensities or states of polarization. These properties, though obtainable from the coefficients of the Fourier development for the classical spectrum, are indeterminate in the quantum theory of Sommerfeld. Bohr has corrected this deficiency partially by the addition of a new postulate to the theory, suggested by the consideration that the requirement mentioned above must be fulfilled not only with regard to the frequency but also to the other properties of the emitted radiation. This postulate, known by the name of *correspondence principle,* is as follows: *The corresponding lines* (in the sense explained above) *have similar intensities and similar states of polarization, and the similarity tends toward identity with increasing quantum numbers.* Hence, calculating the intensities and conditions of polarization of the lines of the classical spectrum by the methods of electromagnetic theory, we obtain an approximate indication concerning the analogous properties of the actually emitted lines. For this reason, the intensity of the line of frequency (354) must, according to the correspondence principle, be approximately proportional to expression (351).

This principle, in spite of its merely qualitative character, has furnished very remarkable results and has been of fundamental importance in the development of the quantum theory, especially since in certain cases (as we shall see in the following section) it has given qualitative as well as quantitative results. We may add that the present theory of radiation, based on quantum mechanics, not only justifies the principle enunciated above but also replaces it by an exact evaluation of the intensity and state of polarization

---

[28] It is apparent that the theorem holds to the same approximation also when the line of the quantum spectrum, of frequency (354), is made to correspond to the line of the classical spectrum having the same indices τ_i but in which the ν_i refer to the final rather than to the initial state (or to any intermediate state). Hence there is a certain arbitrariness in the choice of the correspondence criterion.

of the lines, which is found to be in agreement with experimental results (see §32).

64. The selection principle. One of the most important applications of the correspondence principle obtains in the case in which, in the classical spectrum, the intensities of the lines with indices τ_i are zero, which would be emitted in the state with quantum numbers n_i, in the state of quantum numbers $(n_i - \tau_i)$, or in all intermediate states. Then the quantum line emitted in the transition $n_i \rightarrow n_i - \tau_i$ always corresponds to a classical line of zero intensity, no matter what the criterion adopted to define the correspondence. Hence we may conclude that its intensity is zero, that is, that this quantum jump cannot occur, or is *forbidden.* As may be seen, this is one of the cases in which the correspondence principle leads to a precise—not merely qualitative—statement, which is found to be confirmed by experiment. In this particular case the correspondence principle takes the name of *selection principle.*

We shall add that when in the classical spectrum, in the initial state, in the final state or intermediate states, a single one (or two) of the components X, Y, Z of the electric moment corresponding to a given line is zero, certain conclusions may be drawn concerning the state of polarization of the quantum radiation, as will appear from the following examples.

We shall now see some applications of the selection principle.

(a) *Harmonic oscillator.* The electric moment in this case has only an X-component, given (if the moving particle carries a charge e) by

$$X = ex = eA \sin (2\pi\nu_0 t - \varphi) = \frac{eA}{2i} [e^{i\varphi}\, e^{2\pi i\nu_0 t} - e^{-i\varphi}\, e^{-2\pi i\nu_0 t}].$$

The Fourier expansion therefore reduces to the two terms of frequency ν_0, and hence all the terms whose index τ is not equal to ± 1 are missing. Hence only the transitions in which n varies by ± 1 are possible, and all others are forbidden. We note that, since $E_n = (n + \frac{1}{2})h\nu_0$, the frequency which is emitted (or absorbed) in these transitions is ν_0; that is, the quantum spectrum reduces to a single line whose frequency is exactly equal to that of the classical spectrum or to the proper frequency of the oscillator.

Since the Y- and Z-components of the electric moment are then evidently always zero, the light emitted in the quantum jump will always be polarized along the x-axis, that is, parallel to the direction of oscillation.

(b) *Rigid rotator.* Let us consider a rigid rotator containing electric charges e_i located at distances r_i from the axis. If the axis of rotation is taken along the z-axis, the components of the electric moment are

$$\begin{aligned} X &= \sum_i e_i r_i \cos(2\pi\nu_0 t - \varphi_i), \\ Y &= \sum_i e_i r_i \sin(2\pi\nu_0 t - \varphi_i), \\ Z &= 0, \end{aligned}$$

where ν_0 is the frequency of the rotation and the φ_i are constants. The summations contain as many terms as there are charges. To the first two components we may apply the consideration already mentioned for the oscillator; we find that since only the frequency ν_0 is present in the Fourier development, only those quantum jumps are possible for which n changes by ± 1. Keeping expression (315) for the energy levels in mind, we see that the frequency emitted in the transition $n \rightarrow n - 1$ is

$$\nu_n = (2n + 1)\frac{h}{8\pi^2 I} \tag{355}$$

Therefore the spectrum emitted consists of an infinite number of equidistant lines (on the frequency scale). This result is of considerable importance for the theory of band spectra.

In regard to the state of polarization: since the electric moment rotates in the xy-plane, each line in the classical spectrum would appear circularly polarized if the light is observed along the z-axis, and plane-polarized (parallel to the xy-plane) if the light is observed in a direction perpendicular to the z-axis, whereas elliptically polarized light would be observed along intermediate directions. Since this condition holds good for all the lines of the classical spectrum, we may say as much for the actually emitted lines.

(c) *Selection rule for the azimuthal quantum number.* Let us now apply the selection principle to the central motion of an emission electron, assuming that the field generated by the core is sufficiently close to the Newtonian type that we may decompose the motion into a Keplerian motion plus a uniform precession, as was done in §59. What we shall have to say applies particularly to hydrogenlike systems in which the precession is due solely to a slight relativistic correction.

Let us assume a system of Cartesian coordinates with the x- and y-axes in the (fixed) plane of the orbit. Their relation to the polar coordinates r, ω may be summarized in the formula

$$x + iy = r\, e^{i\omega}. \tag{356}$$

Now we note that because of the precessional motion, r and ω do not have the same period. Indicating by ν_r and ν_ω the respective frequencies, and by ϵ the frequency of the slow precession, we have

$$\nu_\omega = \nu_r + \epsilon. \tag{357}$$

We now introduce a system of axes x', y', which rotate in such a way as to accompany the motion of precession. With respect to these axes the motion will be periodic and hence the angle ω' between the radius vector and the axis x' will be, like r, a periodic function of frequency ν_r, so that we may write the Fourier development

$$r\, e^{i\omega'} = \sum_{\tau=-\infty}^{\infty} A_\tau\, e^{2\pi i \tau \nu_r t}. \tag{358}$$

Since

$$\omega = \omega' + 2\pi\epsilon t,$$

we obtain from (356) and (358),

$$x + iy = r\, e^{i\omega'}\, e^{2\pi i\epsilon t} = \sum_{\tau=-\infty}^{\infty} A_\tau\, e^{2\pi i(\tau\nu_r + \epsilon)t}, \tag{359}$$

or else, solving for ϵ in (357) and putting $\tau_r = \tau - 1$,

$$x + iy = \sum_{\tau_r=-\infty}^{\infty} A_{\tau_r}\, e^{2\pi i(\tau_r\nu_r + \nu_\omega)t}.$$

In order to obtain the expansions of x and y separately, it suffices to write the complex conjugate expression of the last equation and to perform an addition and subtraction. We evidently find two series both containing terms of the type

$$e^{\pm 2\pi i(\tau_r\nu_r + \nu_\omega)t}.$$

We have thus obtained the expansion of the coordinates x and y (and hence of the components of the electric moment) in the form (349). We recognize that whereas τ_r may take on all integral values, the coefficient of ν_ω assumes only the values ± 1. Then, applying the selection principle and recalling that the quantum number related to the coordinate ω is the azimuthal quantum number l, we see that *only those quantum jumps are possible in which the azimuthal quantum number* l *(and hence* k*) changes by* ± 1. We shall express this by writing

$$\Delta l = \pm 1. \tag{360}$$

This is the selection rule for the azimuthal quantum number, which we have already found in §50 by wave mechanics and which is of fundamental importance. In fact, with reference to the term scheme represented in Fig. 45, this rule says that transitions are possible only between adjacent columns of the scheme; this restriction reduces the complication of the spectra enormously. In other words, the s-terms may combine only with p-terms, the p-terms only with s- or d-terms, and so on.

On the other hand, there is no limitation on the changes of the radial quantum number n_r, and hence none on the changes of the principal quantum number n.

It is to be remembered, however, that the preceding reasoning holds under the hypothesis that the motion of the electron differs very little from Keplerian motion, and hence exceptions to the rule are to be expected in the case of terms which are strikingly different from the Balmer terms, and when the atom is subject to strong perturbations (electric or magnetic fields). Actually, in these cases forbidden lines are often observed in the spectrum, but they are of weak intensity, as is to be expected (see footnote 8 on page 226).

(d) *Selection rules for the magnetic quantum number and for the inner quantum number.* An argument similar to the preceding may be made for the magnetic quantum number. For this number to intervene, however, it is necessary to suppose that the atom is situated in a magnetic field, no matter how weak, so that a privileged direction in space is established (see §53), for which we take the z-axis. The magnetic field produces a slow precession about this axis (Larmor theorem). It may then be shown, in a manner similar to (c) above, that for the magnetic quantum number m the selection rule

$$\Delta m = 0, \pm 1$$

holds, which we have already proved in §50 by means of wave mechanics. In addition, we find again the rules for polarization stated in §50.

Finally, for the inner quantum number j we find, by similar considerations, the same selection rule

$$\Delta j = 0, \pm 1,$$

with the additional rule, however, that the transition from $j = 0$ to $j = 0$ is forbidden.

Part III

GENERAL METHODS OF QUANTUM MECHANICS

CHAPTER 10

Mathematical Introduction

1. Function space. It is often convenient to designate an ensemble of N numbers $f_1, f_2, \dots f_N$ by a *point* P in an N-dimensional space (referred to Cartesian axes numbered from 1 to N), or else by a *vector* $\mathbf{f} = OP$ in the same space, calling O the origin of the axes.[1] It is then possible, by an obvious generalization of elementary concepts, to define the *magnitude* or *modulus* $|\mathbf{f}|$ of the vector $\mathbf{f}$ by means of the formula

$$|\mathbf{f}| = f_1^2 + f_2^2 + \cdots + f_N^2;$$

the *scalar product* of two vectors $\mathbf{f}$, $\mathbf{g}$ may be defined by:

$$\mathbf{f} \cdot \mathbf{g} = f_1 g_1 + f_2 g_2 + \cdots + f_N g_N.$$

Also, such operations as sum and difference may be defined in an obvious manner.

If we wish to extend these considerations to the case where $f_1, f_2, \dots f_N$ are complex numbers, it is advantageous to replace the preceding formulas by the following, which, in the case where the numbers are real, reduce to the original formulas (as usual, the asterisk denotes the complex conjugate):

$$|\mathbf{f}| = \sqrt{f_1 f_1^* + f_2 f_2^* + \cdots + f_N f_N^*}, \tag{1}$$

$$\mathbf{f} \cdot \mathbf{g} = f_1 g_1^* + f_2 g_2^* + \cdots + f_N g_N^*. \tag{2}$$

Hence we can say that to specify a vector in the N-dimensional space means to make a (real or complex) number f_r, which is the rth component of the vector, correspond to every integer r (from 1 to N).

We now wish to extend these considerations by introducing a space with an infinite number of dimensions. In order to do this,

---

[1] In this whole chapter we shall deal only with vectors starting at the origin. Therefore to every point there corresponds a vector, and vice versa.

we consider, instead of the N values of the index r from 1 to N, the infinite number of values which a real variable x may take on in an interval (a, b). We can say that instead of N axes, we are considering a (*continuous*) *infinity* of coordinate axes, each corresponding to a value of x. To assign a vector **f** to this space then means to make a (real or complex) number correspond to every value of x (between a and b); that is, it amounts to specifying a function $f(x)$ which is single-valued over (a, b). This operation suggests that we consider any function $f(x)$ which is finite and single-valued in an interval (a, b) (possibly infinite) as a *vector* f in a space with an infinity of dimensions, in which any one of the values of x between a and b characterizes a coordinate axis. The corresponding value taken by the function represents the projection of the vector upon that axis (xth component of the vector).

This space is therefore called *function space.* We can also say that the function $f(x)$ is represented by a *point* in function space; and vice versa, every point of function space represents a function $f(x)$. But often the vectorial interpretation is more useful.

What we have said for a function of *one* variable x may be extended without difficulty to a function of p variables $f(x_1, x_2, \ldots x_p)$, finite and single-valued in a certain domain S (possibly infinite). The function space will in this case have ∞^p dimensions, and every one of its axes will be designated by a group of p numbers. The corresponding component of the vector **f** will be the value which the function f assumes corresponding to these values of the independent variables. In all that follows we shall refer to a function of p variables for greater generality, but we shall often indicate the totality of variables by the single letter x, and shall write $f(x)$ instead of $f(x_1, x_2, \ldots x_p)$.

To the vectors of function space we may immediately apply the usual definitions of sum and difference, as well as that of product of a vector by a scalar.[2] For example, the *sum* of the two vectors **f** and **g** is the vector whose xth component (or, in the case of several variables, the vector whose component along the $x_1, x_2, \ldots x_p$-axis) is $f(x) + g(x)$, that is, the vector representing the function $f(x) + g(x)$. The *product* of the vector **f** by the constant c is the vector c**f** representing the function $cf(x)$.

---

[2] In these considerations, by a scalar is meant a constant quantity (with respect to $x_1, x_2, \ldots x_p$).

If then a certain number n of vectors $\mathbf{f}_1, \mathbf{f}_2, \ldots \mathbf{f}_n$ is given in function space, all the vectors obtainable from the latter by means of linear combinations with arbitrary (constant) coefficients

$$c_1\mathbf{f}_1 + c_2\mathbf{f}_2 + \cdots + c_n\mathbf{f}_n,$$

which represent all the functions expressible as linear combinations of $f_1(x)$, $f_2(x)$ . . . , are said to form a *linear subspace* (*variety, extent*) *in n dimensions.* This is the natural generalization of the concept of plane or of straight line (passing through the origin) in ordinary three-dimensional space, which may be thought of as being specified by means of two vectors or of one vector, respectively, originating from the origin. Of course the same linear subspace may be specified in an infinite number of ways, by selecting for its definition another group of n vectors belonging to it.

There remain to be extended the formulas concerning the magnitude and the scalar product, for which we have to restrict our considerations to a more limited class of functions, as will be shown in the following section.

2. Hilbert space. In order to extend expressions (1), (2), and analogous formulas, to the case with an infinite number of dimensions, we must evidently replace the summations from 1 to N by integrals with respect to x from a to b in the case of a single variable, and by multiple integrals extended over the whole domain S in the general case of p variables. Hence we shall be led to define as *modulus* of the vector $\mathbf{f}$ representing the function f (or, as is sometimes said, as *norm* of the function f) the positive number $|\mathbf{f}|$ whose square is given by the following formula, analogous to (1):

$$|\mathbf{f}|^2 = \int f(x)f^*(x)\,dS. \tag{3}$$

Here and henceforth, the generally multiple integral extended over the whole domain S is indicated by a simple integral sign, and the volume element of the domain by $dS = dx_1\,dx_2 \ldots dx_p$. (In the case of a single variable, $dS = dx$, and the integral is simple.) It is apparent that the integral (3) is not always convergent. Hence we are led to consider from now on *only those functions f for which the integral* (3) *is convergent* (*functions of integrable square*), that is, only for the points of function space for which the distance from the origin has a finite and definite value. The ensemble of these points, called *Hilbert space,* constitutes part of function space. We shall then call *normalized functions*[3] those whose norm is 1 or those represented by a unit vector (or *versor*); any function of integrable square may be normalized by dividing it by its norm.

Let us now extend formula (2) to Hilbert space. The *scalar product* of two vectors **f**, **g** representing the functions $f(x)$, $g(x)$, or *scalar product of the two functions* is the quantity

$$\mathbf{f} \cdot \mathbf{g} = \int f(x) g^*(x)\, dS. \tag{4}$$

(It can be shown that the integral is always convergent because of the convergence of the integrals which define $|\mathbf{f}|$ and $|\mathbf{g}|$.) It is apparent that the scalar product in general is *not* commutative (as it is for real vectors). Reversing the order of its factors changes it into its complex conjugate.

It is also to be noted that if c is a constant,

$$(c\mathbf{f}) \cdot \mathbf{g} = c\mathbf{f} \cdot \mathbf{g}, \tag{5}$$

whereas

$$\mathbf{f} \cdot (c\mathbf{g}) = c^* \mathbf{f} \cdot \mathbf{g}. \tag{5'}$$

Definition (3) of the modulus of a vector **f** or norm of a function f may also be stated as the square root of $\mathbf{f} \cdot \mathbf{f}$.

The *projection* of the vector **f** upon the vector **g** is the quantity

$$f_g = \mathbf{f} \cdot \text{vers}\, \mathbf{g} = f \cdot \frac{\mathbf{g}}{|\mathbf{g}|}. \tag{6}$$

The *condition of orthogonality* of two vectors **f**, **g** [or of the functions $f(x)$, $g(x)$] is that $\mathbf{f} \cdot \mathbf{g} = 0$, because of definition (4), according to which

$$\int f(x) g^*(x)\, dS = 0.$$

This definition of orthogonality between functions has already been introduced in §5 of Part II. Now we see the reason for this designation. (Note that the condition is independent of the order of the two vectors.)

The expansion of a function in a series of orthogonal functions (see §9 of Part II) has an important interpretation in Hilbert space.

---

[3] It is to be noted that this definition coincides with that already given for normalized eigenfunctions in §4 of Part II.

First let us consider the case of a single variable, for simplicity. We observe that any of the orthonormal eigenfunctions $y_n(x)$ (deriving from a differential equation of the type already considered in §3 of Part II:

$$\frac{d}{dx}\left(P\frac{dy}{dx}\right) + Qy = 0, \tag{7}$$

whose eigenvalues we assume to be discrete) is represented by a versor $\mathbf{y}_n$, and that this infinite number of versors are orthogonal to each other. We may therefore say that these versors define, in Hilbert space, a system of an infinity of orthogonal coordinate axes (one for each value of n), in the same way in which a triplet of versors $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ in ordinary space defines[4] a system of Cartesian axes. Now, with the notation (4), relation (32) of Part II may be written

$$f_n = \mathbf{f} \cdot \mathbf{y}_n \tag{8}$$

and is interpreted as follows: the coefficient f_n of the expansion of the function f in terms of the orthogonal functions y_n is the projection of the vector $\mathbf{f}$ upon the versor $\mathbf{y}_n$, or the component of $\mathbf{f}$ along the nth coordinate axis. Then the expansion [(31), Part II], which may be written in vector notation

$$\mathbf{f} = \sum_{n=1}^{\infty} f_n\mathbf{y}_n, \tag{8'}$$

acquires the same meaning as the ordinary-space relation,

$$\mathbf{V} = V_x\mathbf{i} + V_y\mathbf{j} + V_z\mathbf{k},$$

between a vector $\mathbf{V}$ and its components V_x, V_y, V_z.

In Hilbert space, then, we are led to consider, besides the original system of axes corresponding to the infinity of values of x (which we shall call "continuous axes," the values of x constituting a continuous infinity), another system of axes (infinite in number, but discrete), defined by the versors $\mathbf{y}_n$ ($n = 1, 2, \ldots$) and hence depending on the differential equation of which the y_n are eigenfunctions. Every quadratically integrable function $f(x)$ is speci-

---

[4] The origin is understood to be fixed once and for all.

fied[5] by assigning values to the coefficients c_n of its development in a series of eigenfunctions y_n. Consequently, any vector (or any point) of Hilbert space may be specified by means of its components (or its coordinates) relative to the axes defined by the versors $\mathbf{y}_n$, rather than by those components relative to the continuous axes used in the beginning. Thus Hilbert space possesses a *denumerable* infinity of dimensions, not a *continuous* infinity like the function space of which it is a part.

For the axes $\mathbf{y}_n$, the magnitude $|\mathbf{f}|$ of the vector $\mathbf{f}$ may be calculated by means of the formula

$$|\mathbf{f}|^2 = \sum_{n=1}^{\infty} f_n f_n^*, \tag{9}$$

which, by Parseval's formula (see page 99), is equivalent to (3). Similarly, it may easily be shown that the scalar product $\mathbf{f} \cdot \mathbf{g}$ may be calculated by means of the components f_n of $\mathbf{f}$, and g_n of $\mathbf{g}$ by the formula

$$\mathbf{f} \cdot \mathbf{g} = \sum_{n=1}^{\infty} f_n g_n^*, \tag{10}$$

and hence the orthogonality condition may be written

$$\sum_{n=1}^{\infty} f_n g_n^* = 0. \tag{11}$$

Let us now proceed to a consideration of the case of p variables. What has been said above must now be modified in the sense that every eigenfunction is no longer specified by a single index n but rather by p indices, so that we must write $y_{n_1 n_2 \cdots n_p}$. In this way a system of ∞^p orthogonal coordinate axes in Hilbert space is defined, each of which is specified by means of a group of p integers. All the preceding formulas will then have to be modified by inserting

---

[5] Indeed, two functions whose values are equal everywhere, except for some points x constituting an aggregate of zero measure, evidently have the same coefficients c_n. Hence it is useful to consider two such functions as represented by the same vector (or point) of Hilbert space. Moreover, $f(x)$ would be specified (in the sense explained above) by the c_n even if the series $\sum_n c_n y_n$ were not convergent absolutely, but only *in the mean,* in the sense of footnote 5 on page 99.

in the place of n the group $n_1n_2 \ldots n_p$ and by replacing the simple summations by multiple summations. For example, expression (9) becomes

$$|\mathbf{f}|^2 = \sum_{n_1n_2 \cdots n_p \atop 1}^{\infty} f_{n_1n_2 \cdots n_p} f^*_{n_1n_2 \cdots n_p}. \tag{9'}$$

The formulas thus become somewhat more complicated formally, but no new difficulties are introduced. Therefore we shall continue to write the formulas for the case $p = 1$, with the understanding that in order to pass to the general case it will suffice to replace every index by a group of p indices, and each simple summation by a multiple one.

3. Linear operators. We shall call *operator* any symbol which, when written in front of a function (of one or more variables),[6] will change it into another function of the same variables, according to a definite law. Elementary mathematics and the calculus furnish some examples, among which are the following:

(a) Any number k may be regarded as an operator, since when it is put before a $f(x, y, \ldots)$, it changes the latter into the product $kf(x, y, \ldots)$. Of course this convention holds also when k is itself a function. In particular, 1 is an operator which transforms any function into itself; it is called *identity operator.*

(b) The symbols log, sin, cos, and so on, are all operators which change the function $f(x_1, x_2, \ldots)$ into the function $\log f(x_1, x_2, \ldots)$, $\sin f(x_1, x_2, \ldots)$, and so on.

(c) The symbol d/dx is an operator which changes any differentiable function $f(x)$ into its derivative. Similarly (for functions of several variables) the symbols $\partial/\partial x_1$, $\partial/\partial x_2$, and so on, are operators, and also the symbols of second and third derivatives, and so on.

(d) The symbol $\int_{x_0}^{x_1} dx_1$ (with x_0 constant) is an operator which changes any integrable function $f(x_1, x_2, \ldots x_p)$ into the function $\int_{x_0}^{x_1} f(x_1, x_2, \ldots x_p)dx_1$.

---

[6] Sometimes an operator is defined only for certain definite classes of functions and has no meaning for others. For example, the operator d/dx has a meaning only for differentiable functions.

A general operator will be designated by a letter—usually a German letter. For instance, we shall write $F = \mathfrak{A}f$ to indicate that the operator $\mathfrak{A}$ applied to the function f changes it into the function F.

Using geometrical terminology, we may say that an operator defines a correspondence between points (or between vectors) of function space. Therefore we sometimes also write $\mathbf{F} = \mathfrak{A}\mathbf{f}$, for example. We shall always suppose such a one-to-one correspondence.

In what follows we shall deal only with *linear operators*, that is, those which have the two properties:

$$\mathfrak{A}(f + g) = \mathfrak{A}f + \mathfrak{A}g, \qquad (\alpha)$$

where f and g are any two functions;[7] and

$$\mathfrak{A}(cf) = c\mathfrak{A}f, \qquad (\beta)$$

where c is a constant and f is any function. For example, among the operators listed above, the operators k, d/dx, $\partial/\partial x_1$, $\partial/\partial x_2$, $\int_{x_0}^{x_1} dx$, are linear, and the operators log, sin, cos, and so on, are nonlinear.

In function space a linear operator establishes a correspondence between vectors, which is the natural generalization of a *vector mapping* of ordinary three-dimensional space.

The class of linear operators has properties of great mathematical interest (see No. 33 of the Bibliography). We shall confine ourselves here to the essential points of the theory of these operators.

4. Linear operator algebra. We can define operations of combination among linear operators, analogous to the operations sum, difference, and so on, by which algebraic quantities may be combined with each other. Such definition permits the construction of a linear operator algebra analogous (although not identical) to ordinary algebra.

Given two linear operators $\mathfrak{A}$ and $\mathfrak{B}$, we call their *sum*, and indicate by $\mathfrak{A} + \mathfrak{B}$, the (linear) operator defined by[8]

$$(\mathfrak{A} + \mathfrak{B})f = \mathfrak{A}f + \mathfrak{B}f.$$

---

[7] Of course, provided that they are such that it is meaningful to apply the operator $\mathfrak{A}$ to them. This condition will always be implied in what follows.

[8] Here and in what follows, f is any function to which the operators in question may be applied.

In an analogous manner we define the *difference* of two linear operators, and the sum of any number of them.

For instance, take the linear operator Δ, called the *Laplacian*, which is used a great deal in mathematical physics. It is defined by

$$\Delta f(x, y, z) = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2},$$

the sum of the three linear operators $\partial^2/\partial x^2$, $\partial^2/\partial y^2$, $\partial^2/\partial z^2$; that is, we may write

$$\Delta \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}. \tag{12}$$

Another well-known example is the operator which occurs in the left-hand member of the Schrödinger equation (131) of Part II, and may be written

$$\left[\Delta + \frac{8\pi^2 m}{h^2}(E - U)\right]\psi = 0. \tag{13}$$

The operation of first applying the operator $\mathfrak{B}$ and then applying, upon the function obtained, the operation $\mathfrak{A}$, is called the *product* of the linear operator $\mathfrak{A}$ by the linear operator $\mathfrak{B}$, and is indicated by $\mathfrak{A}\mathfrak{B}$. This amounts to saying that

$$\mathfrak{A}\mathfrak{B}f = \mathfrak{A}(\mathfrak{B}f).$$

Evidently the product of two linear operators is itself linear.

In general, the commutative law does *not* hold for such a product; that is, the operator $\mathfrak{A}\mathfrak{B}$ does not coincide with the operator $\mathfrak{B}\mathfrak{A}$. It is just this property which makes operator algebra different from ordinary algebra. When the two linear operators $\mathfrak{A}\mathfrak{B}$ and $\mathfrak{B}\mathfrak{A}$ are identical, $\mathfrak{A}$ and $\mathfrak{B}$ are said to *commute.*

Examples. Two numerical factors k_1, k_2 (either constants or not) are always commuting operators, since $k_1k_2f = k_2k_1f$. Similarly, the linear operators $\partial/\partial x$ and $\partial/\partial y$ commute as a rule, if their product in the order indicated is $(\partial^2/(\partial x\,\partial y)$ and in reversed order is $\partial^2/\partial y\,\partial x$. However, the two linear operators

$$\mathfrak{A} = x \qquad \text{and} \qquad \mathfrak{B} = \frac{\partial}{\partial x} \tag{14}$$

do *not* commute, because

$$\mathfrak{A}\mathfrak{B}f = x\frac{\partial f}{\partial x}$$

$$\mathfrak{B}\mathfrak{A}f = \frac{\partial}{\partial x}(xf) = f + x\frac{\partial f}{\partial x}.$$

The last equality may also be written

$$\mathfrak{B}\mathfrak{A}f = f + \mathfrak{A}\mathfrak{B}f = (1 + \mathfrak{A}\mathfrak{B})f,$$

which means that for the operators $\mathfrak{A}$ and $\mathfrak{B}$ defined by (14), the *commutation relation*

$$\mathfrak{B}\mathfrak{A} - \mathfrak{A}\mathfrak{B} = 1 \tag{15}$$

holds, instead of the commutative property.

Obviously, the nth *power* (with n a positive integer) of a linear operator $\mathfrak{A}$ is defined as the product of n factors equal to $\mathfrak{A}$. We also adopt the convention $\mathfrak{A}^0 = 1$.

Example. The nth power of the operator $\frac{\partial}{\partial x}$ is $\frac{\partial^n}{\partial x^n}$; that is,

$$\left(\frac{\partial}{\partial x}\right)^n \equiv \frac{\partial^n}{\partial x^n}.$$

Hence (12) may also be written

$$\Delta \equiv \left(\frac{\partial}{\partial x}\right)^2 + \left(\frac{\partial}{\partial y}\right)^2 + \left(\frac{\partial}{\partial z}\right)^2.$$

Given a linear operator $\mathfrak{A}$, if there exists a linear operator $\mathfrak{B}$ such that

$$\mathfrak{A}\mathfrak{B} = \mathfrak{B}\mathfrak{A} = 1,$$

we say that $\mathfrak{B}$ is the *reciprocal* or the *inverse* of $\mathfrak{A}$, and vice versa, and we write

$$\mathfrak{B} = \mathfrak{A}^{-1}, \qquad \mathfrak{A} = \mathfrak{B}^{-1}.$$

If we apply to a function an operator and its inverse in succession, the two operations will cancel, and again we have the original function.

If the inverse of $\mathfrak{A}$ exists, we may define the powers of $\mathfrak{A}$ with negative exponent, putting

$$\mathfrak{A}^{-n} = (\mathfrak{A}^{-1})^n. \tag{16}$$

It is clear that for operators the ordinary theorems on powers hold; for example, $\mathfrak{A}^n\mathfrak{A}^m = \mathfrak{A}^{n+m}$ (n, m are positive, zero, or negative), and so on.

It is apparent that if both $\mathfrak{A}$ and $\mathfrak{B}$ possess an inverse, their product $\mathfrak{A}\mathfrak{B}$ will also have one, and it will be $\mathfrak{B}^{-1}\mathfrak{A}^{-1}$ (note the inversion of factors). In fact,

$$(\mathfrak{A}\mathfrak{B})(\mathfrak{B}^{-1}\mathfrak{A}^{-1}) = \mathfrak{A}(\mathfrak{B}\mathfrak{B}^{-1})\mathfrak{A}^{-1} = \mathfrak{A}\mathfrak{A}^{-1} = 1$$

The powers of a linear operator $\mathfrak{A}$ having been defined, we may define other linear operators called *functions* of $\mathfrak{A}$, in the following manner. Let $F(\alpha)$ be the symbol of a function which can be expanded in a power series in the variable α; that is, within a certain circle of convergence, let

$$F(\alpha) = a_0 + a_1\alpha + a_2\alpha^2 + \cdots .$$

We shall then define as $F(\mathfrak{A})$ the linear operator obtained by replacing bodily in the preceding series the symbol α by the symbol $\mathfrak{A}$. Every term of the series will then acquire the meaning of a well-defined operator.

Example. Let us take for $\mathfrak{A}$ the linear operator $\lambda(\partial/\partial x)$ (where λ is a constant), and let us define the linear operator $e^{\mathfrak{A}}$ or $e^{\lambda(\partial/\partial x)}$. Since the function e^α is defined by the series

$$e^\alpha = 1 + \alpha + \frac{1}{2!}\alpha^2 + \frac{1}{3!}\alpha^3 + \cdots ,$$

we have, replacing α by $\lambda(\partial/\partial x)$,

$$e^{\lambda(\partial/\partial x)} = 1 + \lambda\frac{\partial}{\partial x} + \frac{\lambda^2}{2!}\frac{\partial^2}{\partial x^2} + \frac{\lambda^3}{3!}\frac{\partial^3}{\partial x^3} + \cdots$$

The operator on the right has a rather important interpretation: it changes $f(x)$ into $f(x + \lambda)$, since, by the Taylor formula,

$$f + \lambda\frac{\partial f}{\partial x} + \frac{\lambda^2}{2!}\frac{\partial^2 f}{\partial x^2} + \frac{\lambda^3}{3!}\frac{\partial^3 f}{\partial x^3} + \cdots = f(x + \lambda).$$

Hence we shall write

$$e^{\lambda(\partial/\partial x)}f(x) = f(x + \lambda).$$

It is evident that a linear operator $\mathfrak{A}$ commutes with any power $\mathfrak{A}^n$, and hence also with any $F(\mathfrak{A})$.

One further sees immediately that if an operator $\mathfrak{B}$ commutes with $\mathfrak{A}$, it also commutes with any $F(\mathfrak{A})$.

We now proceed to define a function of several linear operators $\mathfrak{A}$, $\mathfrak{B}$, $\mathfrak{C}$, confining ourselves (for ease of writing) to the case of two. Given a function of two variables which may be expanded

$$F(\alpha, \beta) = \sum_{i,k} a_{ik}\alpha^i\beta^k, \tag{17}$$

the immediate extension of the last case leads us to define $F(\mathfrak{A}, \mathfrak{B})$ by the linear operator

$$\mathfrak{F} = \sum_{ik} a_{ik}\mathfrak{A}^i\mathfrak{B}^k. \tag{18}$$

However, it must be pointed out that although this definition may be adopted without hesitation if $\mathfrak{A}$ and $\mathfrak{B}$ commute, the case where they do not commute requires some attention. In fact, in a general term of (17) the order of the factors α and β is indifferent, so that instead of $\alpha^i\beta^k$ we could write $\alpha^{i-1}\beta^k\alpha$ or $\alpha^{i-2}\beta^k\alpha^2$, for example, or many other forms. The same function $F(\alpha, \beta)$ always corresponds to all these ways of writing. On the other hand, if we replace α and β by two noncommuting linear operators $\mathfrak{A}$ and $\mathfrak{B}$, we obtain as many different linear operators from these series. Therefore it is necessary to make the linear operator $\mathfrak{F} = F(\mathfrak{A}, \mathfrak{B})$ correspond, not to the analytic function $F(\alpha, \beta)$, but to a *particular way* of writing that function.

We then immediately realize that a linear operator which is a function of one or more commuting linear operators $\mathfrak{A}$, $\mathfrak{B}$, . . . will itself commute with each of these operators.

5. Representation of a linear operator by a matrix. It is easy to show that if $\mathfrak{A}$ is a linear operator which operates upon vectors of Hilbert space, then the components of the vector $\mathfrak{A}\mathbf{f}$ (which we shall indicate by $\mathbf{F}$) are linear combinations of the components of $\mathbf{f}$. This is equivalent to saying that any linear operator is equivalent to a linear transformation of the components of the vector to which it is applied (analogously to what occurs in vector mappings in ordinary space).

In fact, if as usual we call $\mathbf{y}_n$ the versors of the axes (eigenfunctions of a differential equation), we may write each vector $\mathbf{f}$ in the form

$$\mathbf{f} = \sum_n f_n\mathbf{y}_n, \tag{19}$$

where the f_n are the components of $\mathbf{f}$, and the sum is understood to be extended from 1 to ∞ (this will also be implied for the formulas which follow). Applying the operator $\mathfrak{A}$, since it is linear, we have

$$\mathbf{F} = \mathfrak{A}\mathbf{f} = \sum f_n\mathfrak{A}\mathbf{y}_n \tag{20}$$

so that it is sufficient to know the effect of the linear operator $\mathfrak{A}$ on the fundamental versors $\mathbf{y}_n$ in order to be able to apply it to any $\mathbf{f}$. Now, the vector $\mathfrak{A}\mathbf{y}_n$ will be specified by its components which we shall indicate by A_{mn} (where the first subscript characterizes the component and the second specifies the vector with which we are dealing). Hence we may write

$$\mathfrak{A}\mathbf{y}_n = \sum_m A_{mn}\mathbf{y}_m. \tag{21}$$

Substituting into (20), we have

$$\mathbf{F} = \sum_n f_n A_{mn}\mathbf{y}_m,$$

or else, writing

$$F_m = \sum_n A_{mn} f_n, \tag{22}$$

we have

$$\mathbf{F} = \sum_m F_m \mathbf{y}_m. \tag{20'}$$

From this sequence we see that the components of the vector $\mathfrak{A}$ are the F_m, which are obtained from the components f_n of $\mathbf{f}$ by means of the system of (infinitely many) linear relations (22), whose coefficients are the A_{mn}. Hence a knowledge of these coefficients permits us to obtain for any $\mathbf{f}$ the corresponding $\mathfrak{A}\mathbf{f}$, and the ensemble of these coefficients completely determines the operator $\mathfrak{A}$.

The coefficients A_{mn} constitute a double (denumerable) infinity of (generally complex) numbers, which may be placed in an array (*matrix*) of an infinite number of rows, characterized by the first subscript, and of an infinite number of columns, specified by the second subscript:

$$\left\| \begin{array}{cccc} A_{11} & A_{12} & A_{13} & \cdots \\ A_{21} & A_{22} & A_{23} & \cdots \\ A_{31} & A_{32} & A_{33} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{array} \right\|.$$

Thus to every linear operator $\mathfrak{A}$ there corresponds a matrix which perfectly specifies it and which is generally indicated by the same symbol $\mathfrak{A}$ as the operator. Often, when it is necessary to emphasize that we are dealing with the matrix, we shall designate it by $\{\mathfrak{A}\}$.

The *elements* A_{mn} of this matrix may be calculated, with the observation that A_{mn} represents the mth component of the vector $\mathfrak{A}\mathbf{y}_n$, and hence [see formula (8)]

$$A_{mn} = (\mathfrak{A}\mathbf{y}_n) \cdot \mathbf{y}_m. \tag{23}$$

Or, recalling (4) and writing the factor with the asterisk first to avoid ambiguities, we have

$$A_{mn} = \int y_m^* \mathfrak{A} y_n \, dx. \tag{23'}$$

Example. Let $\mathfrak{A}$ be the identity operator ($\mathfrak{A} = 1$). Then (23) yields

$$A_{mn} = \int y_m^* y_n \, dx = \delta_{mn}, \tag{24}$$

and hence the matrix representing the identity operator is

$$\begin{Vmatrix} 1 & 0 & 0 & \ldots \\ 0 & 1 & 0 & \ldots \\ 0 & 0 & 1 & \ldots \\ \ldots & \ldots & \ldots & \ldots \end{Vmatrix}. \tag{25}$$

This is called the *unit matrix* and is indicated by $\{1\}$.

Note. In the general case of p variables, every element of the matrix must be written in the form $A_{m_1m_2 \cdots m_pn_1n_2 \cdots n_p}$, for reasons mentioned in §2; that is, we shall be dealing with a matrix in which the rows and columns are characterized, not by two indices, but by two groups of p indices. This situation does not introduce any conceptual difficulty but merely complicates the writing considerably.

6. Matrix algebra. Since to any linear operator (fixed in the system of reference) there corresponds a matrix, and vice versa, it is evident that by the operations of sum, difference, and so on, defined between linear operators, as many operations between matrices are established. For instance, we shall call *sum* of the matrices $\{\mathfrak{A}\}$ and $\{\mathfrak{B}\}$, and shall indicate by $\{\mathfrak{A}\} + \{\mathfrak{B}\}$ the matrix $\{\mathfrak{A} + \mathfrak{B}\}$—that is, the matrix corresponding to the linear operator which is the sum of $\mathfrak{A}$ and $\mathfrak{B}$. The *product* of two matrices will be defined in a similar way, from which also the nth *power* (n being a positive integer) of a matrix can be derived. Similarly, we shall define any *analytic function* $F(\{\mathfrak{A}\})$ of a matrix as the matrix corresponding to the linear operator $F(\mathfrak{A})$, defined in §4; and analogously for a function of several matrices. If $\mathfrak{A}$ possesses an inverse $\mathfrak{A}^{-1}$, the matrices $\{\mathfrak{A}\}$ and $\{\mathfrak{A}^{-1}\}$ will be called *inverse*, or reciprocal of each other. Their product (in any order) is the unit matrix (25). The matrix $\{\mathfrak{A}^{-1}\}$ will be indicated by $\{\mathfrak{A}\}^{-1}$, and its powers will be considered as powers of $\{\mathfrak{A}\}$ with negative exponents.

It is preferable to give a definition which is equivalent to these operations between matrices but which is independent of the respective operators; consequently, we shall define the operations between matrices by means of operations to be performed upon their elements. This is done as follows.

The elements of the *sum* matrix $\{\mathfrak{A}\} + \{\mathfrak{B}\}$ are the sums of corresponding elements of the two matrices $\{\mathfrak{A}\}$ and $\{\mathfrak{B}\}$; that is,

$$\{\mathfrak{A}\} + \{\mathfrak{B}\} = \left\| \begin{matrix} A_{11}+B_{11} & A_{12}+B_{12} & A_{13}+B_{13} & \cdots \\ A_{21}+B_{21} & A_{22}+B_{22} & A_{23}+B_{23} & \cdots \\ A_{31}+B_{31} & A_{32}+B_{32} & A_{33}+B_{33} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{matrix} \right\|.$$

From (23) it follows immediately that the matrix sum, defined in this way, is actually the matrix corresponding to the operator $\mathfrak{A} + \mathfrak{B}$.

The same is true for the *difference* of two matrices and for the sum of any number of matrices.

The *multiplication* of a matrix $\{\mathfrak{A}\}$ by a constant k is performed by multiplying each element of the matrix by k. This statement also follows immediately from (23).

Let us proceed to the *product* of two matrices $\mathfrak{A}$, $\mathfrak{B}$. Let us call $\mathfrak{P}$ the linear operator

$$\mathfrak{P} = \mathfrak{A}\mathfrak{B} \tag{26}$$

and let us calculate by means of (23) the general element of the matrix product $\{\mathfrak{P}\} = \{\mathfrak{A}\}\{\mathfrak{B}\}$:

$$P_{mn} = (\mathfrak{P}\mathbf{y}_n) \cdot \mathbf{y}_m = (\mathfrak{A}\mathfrak{B}\mathbf{y}_n) \cdot \mathbf{y}_m. \tag{27}$$

Let us calculate the factor in parentheses. Because of (21), we have

$$\mathfrak{B}\mathbf{y}_n = \sum_i B_{in}\mathbf{y}_i$$

and hence

$$\mathfrak{A}\mathfrak{B}\mathbf{y}_n = \sum_i B_{in}\mathfrak{A}\mathbf{y}_i = \sum_i B_{in} \sum_k A_{ki}\mathbf{y}_k = \sum_k \Big(\sum_i A_{ki}B_{in}\Big) \mathbf{y}_k.$$

Substituting into (27) and recalling that the y are orthonormal, we have

$$P_{mn} = \sum_i A_{mi} B_{in}. \tag{28}$$

As a memory aid, note that by writing, as we have done, the two factors A_{mi} and B_{in} in the order corresponding to that of (26), the two subscripts m and n occur in the same order in the two terms, and the index of summation remains between them.

Formula (28) may readily be seen to be equivalent to the following rule: The product of two matrices is obtained by the known rule for the product of two determinants, but always by multiplying the rows of the first matrix by the columns of the second matrix, and not conversely.

Of course the product of two matrices is *not* commutative, except for the case in which the two corresponding operators commute, whereupon the two matrices are said to *commute.*

By applying the preceding rule to the unit matrix (25) and to any other matrix $\{\mathfrak{A}\}$, it may readily be verified that

$$\{1\}\{\mathfrak{A}\} = \{\mathfrak{A}\}\{1\} = \{\mathfrak{A}\}.$$

6a. Representation of a function by means of a matrix with a single column. We may give an expressive interpretation to the formulas of §5 if we agree to consider the components f_n (which characterize a function f with respect to a certain system of orthogonal functions) to be elements of a matrix with a single column (and an infinity of rows). That is, we write

$$f = \begin{vmatrix} f_1 \\ f_2 \\ f_3 \\ \cdot \\ \cdot \\ \cdot \end{vmatrix}.$$

Now not only operators but also functions are represented by matrices, and we may readily verify that in order to operate on a function f with an operator $\mathfrak{A}$ it is sufficient to form the product of the corresponding matrices by the ordinary rule, that is, by multiplying "rows by columns." In fact, upon applying this rule and keeping (22) in mind, we readily find

$$\begin{vmatrix} A_{11} & A_{12} & A_{13} & \cdots \\ A_{21} & A_{22} & A_{23} & \cdots \\ A_{31} & A_{32} & A_{33} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{vmatrix} \cdot \begin{vmatrix} f_1 \\ f_2 \\ f_3 \\ \cdot \end{vmatrix} = \begin{vmatrix} F_1 \\ F_2 \\ F_3 \\ \cdot \end{vmatrix}. \tag{29}$$

7. Change of axes in Hilbert space. Transformation matrices. Let us now consider, in addition to the eigenfunctions y_n defined by equation (7), another complete system of eigenfunctions y'_n[9] defined by another equation of the same form (referring to the same interval and with the same boundary conditions). This new set will define another system of orthogonal axes, specified by the versors y'_n. With respect to these axes the vector **f** will have certain components f'_n expressed, just as for f_n, by

$$f'_n = \mathbf{f} \cdot \mathbf{y}'_n, \tag{30}$$

and we may write, in analogy to (8′),

$$\mathbf{f} = \sum_n f'_n \mathbf{y}'_n. \tag{31}$$

Now let us look for the relation between the components of the vector **f** in the new and the old axes, that is, between f'_n and f_n We note that each of the versors $\mathbf{y}'_n$ may be specified by its components (with respect to the old axes) which we shall designate by S_{mn} (the first subscript specifies the component, the second subscript the versor with which we are dealing), so that

$$\mathbf{y}'_n = \sum_m S_{mn} \mathbf{y}_m. \tag{32}$$

Because of (8), we have

$$S_{mn} = \mathbf{y}'_n \cdot \mathbf{y}_m \tag{33}$$

and the quantities S_{mn} will obey the relations [see formula (10)]:

$$\sum_i S_{im} S^*_{in} = \delta_{mn}, \tag{34}$$

which express the fact that the y'_n are orthonormal.

---

[9] It need hardly be pointed out that the prime here (and in this whole section) does not have the meaning "derivative."

With this and from (30) and (32), using (5), we obtain for the new components the following expression:

$$f'_n = \sum_m S^*_{mn} \mathbf{f} \cdot \mathbf{y}_m,$$

or, because of (8),

$$f'_n = \sum_m S^*_{mn} f_m. \tag{35}$$

In order to obtain the inverse formulas, we could solve these for the f_m, but it is more convenient to proceed in symmetrical fashion, that is, to consider the components of the versors $\mathbf{y}$ with respect to the axes $\mathbf{y}'$. We shall indicate these components by $\tilde{S}_{mn}$, putting

$$\tilde{S}_{mn} = \mathbf{y}_n \cdot \mathbf{y}'_m, \tag{36}$$

so that

$$\mathbf{y}_n = \sum_m \tilde{S}_{mn} \mathbf{y}'_m. \tag{32'}$$

We then find, analogously to (35),

$$f_n = \sum_m \tilde{S}^*_{mn} f'_m. \tag{35'}$$

We then note that

$$\tilde{S}_{mn} = (\mathbf{y}'_m \cdot \mathbf{y}_n)^* = S^*_{mn}, \tag{37}$$

so that we may also write

$$f_n = \sum_m S_{mn} f'_m. \tag{35''}$$

It is convenient to consider the S_{mn} as elements of a matrix $\{\mathfrak{S}\}$. We shall say that we may pass from the components of a vector with respect to the y-axes to its components with respect to the y'-axes by means of the matrix $\{\mathfrak{S}\}$—that is, by means of the linear transformation (32). Similarly, the matrix $\{\tilde{\mathfrak{S}}\}$ effects the passage from the components y' to the components y by means of the transformation (32'). Furthermore, (37) expresses the fact that the matrix $\{\tilde{\mathfrak{S}}\}$ is obtained from $\{\mathfrak{S}\}$ by interchanging its rows and columns (that is, by replacing each element with the one symmetrical with respect to the principal diagonal) and taking the complex conjugate of each element. If we form the product of these two matrices $(\{\tilde{\mathfrak{S}}\}\{\mathfrak{S}\})$ by the rule of §6, we find, for its general element with indices m and n, meanwhile recalling (37) and (34),

$$\sum_i S_{mi} S_{in} = \sum_i S^*_{im} S_{in} = \delta_{mn}.$$

Hence we can write

$$\{\tilde{\mathfrak{S}}\}\{\mathfrak{S}\} = \{1\}.$$

Similarly, it can be proved that $\{\mathfrak{S}\}\{\tilde{\mathfrak{S}}\} = 1$. Hence the matrix $\{\tilde{\mathfrak{S}}\}$ is the *inverse* of $\{\mathfrak{S}\}$:

$$\{\tilde{\mathfrak{S}}\} = \{\mathfrak{S}\}^{-1}. \tag{38}$$

This property, equivalent to (34), characterizes matrices which are said to be *unitary*.

The matrices introduced in this way are not considered to represent operators, since they are not used to pass from one vector to another, but instead from the components of a vector with respect to one system of reference to the components of the same vector with respect to another system of reference. They are therefore called *transformation matrices.*

Let us suppose that after having passed from the representation y to the representation y' by means of the matrix $\{\mathfrak{S}\}$, we pass on to a third (complete and orthogonal) representation of versors $\mathbf{y}''$, by means of another transformation matrix $\{\mathfrak{T}\}$. We can then show that it is possible to pass directly from the representation y to the representation y'' by means of the transformation matrix

$$\{\mathfrak{P}\} = \{\mathfrak{S}\}\{\mathfrak{T}\}.$$

In fact, the passage from f to f'' will be expressed by the following formula, analogous to (35):

$$f_l'' = \sum_n T^*_{nl} f_n'. \tag{39}$$

Substituting (35) for f_n', we have

$$f_l'' = \sum_{m,n} T^*_{nl} S^*_{mn} f_m = \sum_m \left(\sum_n S^*_{mn} T^*_{nl} \right) f_m$$

and. by making use of the rule for matrix multiplication (28),

$$f_l'' = \sum_m P^*_{ml} f_m. \tag{40}$$

Upon comparison with (35), this operation shows that we may pass from the f to the f'' by means of the matrix $\{\mathfrak{P}\}$ in the same way in which the matrix $\{\mathfrak{S}\}$ effects the passage from the f to the f'.

8. Representation of the same operator in different systems of reference. We have seen that an operator $\mathfrak{A}$, fixed in a system of axes (specified by the versors $\mathbf{y}_n$) in Hilbert space, is represented by a matrix $\{\mathfrak{A}\}$ whose elements are given by (20). If one now considers another system of axes $\mathbf{y}'_n$, as we did in the last section, the problem arises of finding the matrix $\{\mathfrak{A}\}'$ which represents the same operator in the new system of reference, that is, the matrix whose elements A'_{mn} express the components F'_m of $\mathfrak{A}\mathbf{f}$ (with respect to the axes $\mathbf{y}'_n$) in terms of the components f'_n of $\mathbf{f}$ by the formula

$$F'_m = \sum_n A'_{mn} f'_n, \tag{41}$$

which corresponds to (22).

The element A'_{mn} will be given by the formula

$$A'_{mn} = (\mathfrak{A}\mathbf{y}'_n) \cdot \mathbf{y}'_m, \tag{42}$$

which corresponds to (23).

Now we substitute for $\mathbf{y}'_n$ and $\mathbf{y}'_m$ their expressions in terms of $\mathbf{y}$, namely [see (32)],

$$\mathbf{y}'_n = \sum_i S_{in}\mathbf{y}_i, \qquad \mathbf{y}'_m = \sum_k S_{km}\mathbf{y}_k,$$

and we get, recalling (23),

$$A'_{mn} = \sum_{i,k} S^*_{km} A_{ki} S_{in}, \tag{43}$$

or else, because of (37),

$$A'_{mn} = \sum_{i,k} \tilde{S}_{mk} A_{ki} S_{in}.$$

However, by the multiplication rule, this is simply the element (m, n) of the matrix $\{\tilde{\mathfrak{S}}\}\{\mathfrak{A}\}\{\mathfrak{S}\}$, or else, by (38), $\{\mathfrak{S}\}^{-1}\{\mathfrak{A}\}\{\mathfrak{S}\}$.

Hence we write

$$\{\mathfrak{A}\}' = \{\mathfrak{S}\}^{-1}\{\mathfrak{A}\}\{\mathfrak{S}\} \tag{44}$$

This is the law according to which the matrix $\{\mathfrak{A}\}$ transforms in going from the y- to the y'-axes.

It need hardly be mentioned that the algebraic relations between matrices have the same form in any system of reference. If, for instance, we have in the first system

$$\{\mathfrak{A}\} = F(\{\mathfrak{B}\}) \tag{45}$$

(F standing for an analytic function), we still have in the second system

$$\{\mathfrak{A}\}' = F(\{\mathfrak{B}\}'). \tag{45'}$$

This is a direct consequence of the fact that both (45) and (45′) express the same relation $\mathfrak{A} = F(\mathfrak{B})$ between the operators. Furthermore, it would be easy to prove this directly by verifying that the transformation (44) does not alter the relations for sum, product, and inverse, from which any analytic relation may be constructed.

9. Hermitian operators and matrices. Of particular interest in quantum mechanics are the linear operators $\mathfrak{A}$ possessing the property that for any function f the product $\mathbf{f} \cdot \mathfrak{A}\mathbf{f}$ is real, that is,

$$\mathbf{f} \cdot \mathfrak{A}\mathbf{f} = (\mathbf{f} \cdot \mathfrak{A}\mathbf{f})^*. \tag{46}$$

These operators are called *Hermitian*.

As an important example, let us consider the operator which occurred in §1 of Part II, that is, (A, B, C being real),

$$\mathfrak{L} = A\frac{d^2}{dx^2} + B\frac{d}{dx} + C, \tag{47}$$

and let us look for the condition which makes $\mathfrak{L}$ Hermitian. Applying (46), we see that for any f we must have

$$\int A(f''f^* - f^{*\prime\prime}f)dx + \int B(f'f^* - f^{*\prime}f)dx = 0.$$

Noting that the first of the two differences in parentheses is the derivative of the second one, and calculating the first integral by integration by parts, we obtain (if f vanishes at the endpoints)

$$\int(A' - B)(f'f^* - f^{*\prime}f)dx = 0,$$

which requires, since f is arbitrary, that $A' = B$. Hence we find the condition which we have already stated by saying that the equation $\mathfrak{L}f = 0$ is *self-adjoint* (see §3 of Part II).

An operator $\mathfrak{A}$'s property of being Hermitian may be carried over into an important property of the matrix representing it (in any system of axes). In fact, it is apparent that if, as usual, we set $F = \mathfrak{A}f$, we have, because of (10) and (22),

$$\mathbf{f} \cdot \mathfrak{A}\mathbf{f} = \sum_m f_m F_m^* = \sum_{m,n} f_m A_{mn}^* f_n^*, \tag{48}$$

and taking its complex conjugate,

$$(\mathbf{f} \cdot \mathfrak{A}\mathbf{f})^* = \sum_{m,n} f_m^* A_{mn} f_n.$$

This equation may also be written, by interchanging the summation indices, as

$$(\mathbf{f} \cdot \mathfrak{A}\mathbf{f})^* = \sum_{m,n} f_n^* A_{nm} f_m. \tag{48'}$$

Substituting the last expression, along with (48), into the condition for Hermitian matrices (46), we obtain

$$A_{mn}^* = A_{nm}, \tag{46'}$$

that is, the elements lying symmetrically with respect to the principal diagonal are complex conjugates of each other. (In particular, the elements along the principal diagonal will be real.) With the notation explained in §7, (46') may be written $\{\tilde{\mathfrak{A}}\} = \{\mathfrak{A}\}$. Such a matrix is called *Hermitian.* It is obvious that, conversely, a Hermitian matrix always represents a Hermitian operator.

From this analysis we may easily obtain another property of Hermitian operators: for any two functions f and g if (and only if) $\mathfrak{A}$ is Hermitian, we have

$$\mathbf{f} \cdot \mathfrak{A}\mathbf{g} = \mathfrak{A}\mathbf{f} \cdot \mathbf{g}. \tag{49}$$

In fact, calling g_n the components of $\mathbf{g}$, we have, analogously to (48),

$$\mathbf{f} \cdot \mathfrak{A}\mathbf{g} = \sum_{m,n} f_m A_{mn}^* g_n^*,$$

$$\mathbf{g} \cdot \mathfrak{A}\mathbf{f} = \sum_{m,n} g_m A_{mn}^* f_n^*.$$

Taking the complex conjugate of the second summation and interchanging the indices m and n, we recognize that, by virtue of (46′), this summation is identical with the first one; that is, (49) is verified.

It is apparent that if $\{\mathfrak{A}\}$ is a Hermitian matrix, so is the matrix $\{\mathfrak{A}\}'$ which corresponds to the former in any other system of reference. This fact may be recognized either by means of (44) or by observing that if $\{\mathfrak{A}\}$ is Hermitian, the operator which it represents will also be Hermitian, and hence the operator will be represented by a Hermitian matrix even in another reference frame.

In what follows we shall deal only with Hermitian operators and matrices.

We shall now show that if $\mathfrak{A}$ and $\mathfrak{B}$ are linear Hermitian operators, then the two linear operators

$$\mathfrak{C} = \mathfrak{A}\mathfrak{B} + \mathfrak{B}\mathfrak{A} \tag{50}$$

and

$$\mathfrak{D} = i(\mathfrak{A}\mathfrak{B} - \mathfrak{B}\mathfrak{A}) \tag{50′}$$

are Hermitian, and in particular that if $\mathfrak{A}$ and $\mathfrak{B}$ commute, their product is Hermitian.

Indeed, from (49) we have, replacing **g** (an arbitrary vector) by $\mathfrak{B}\mathbf{g}$,

$$\mathbf{f} \cdot \mathfrak{A}\mathfrak{B}\mathbf{g} = \mathfrak{A}\mathbf{f} \cdot \mathfrak{B}\mathbf{g}. \tag{51}$$

We shall similarly find the formula [which may be obtained from (51) by interchanging $\mathfrak{A}$ and $\mathfrak{B}$, and **f** and **g**]

$$\mathbf{g} \cdot \mathfrak{B}\mathfrak{A}\mathbf{f} = \mathfrak{B}\mathbf{g} \cdot \mathfrak{A}\mathbf{f}$$

or else

$$\mathfrak{B}\mathfrak{A}\mathbf{f} \cdot \mathbf{g} = \mathfrak{A}\mathbf{f} \cdot \mathfrak{B}\mathbf{g}.$$

A comparison with (51) shows that

$$\mathbf{f} \cdot \mathfrak{A}\mathfrak{B}\mathbf{g} = \mathfrak{B}\mathfrak{A}\mathbf{f} \cdot \mathbf{g}.$$

Similarly, interchanging $\mathfrak{A}$ and $\mathfrak{B}$,

$$\mathbf{f} \cdot \mathfrak{B}\mathfrak{A}\mathbf{g} = \mathfrak{A}\mathfrak{B}\mathbf{f} \cdot \mathbf{g}.$$

Adding the last two expressions and using definition (50), we obtain

$$\mathbf{f} \cdot \mathfrak{C}\mathbf{g} = \mathfrak{C}\mathbf{f} \cdot \mathbf{g}$$

which expresses the fact that $\mathfrak{C}$ is Hermitian. Subtracting the two expressions and using definition (50′), we find

$$\mathbf{f} \cdot i\mathfrak{D}\mathbf{g} = -i\mathfrak{D}\mathbf{f} \cdot \mathbf{g}; \tag{51′}$$

and since $\mathbf{f} \cdot i\mathfrak{D}\mathbf{g} = -i\mathbf{f} \cdot \mathfrak{D}\mathbf{g}$ because of (5′), (51′) may be written

$$\mathbf{f} \cdot \mathfrak{D}\mathbf{g} = \mathfrak{D}\mathbf{f} . \mathbf{g}$$

expressing the fact that $\mathfrak{D}$ is Hermitian.

A corollary of the preceding theorem states that if $\mathfrak{A}$ is Hermitian, then all its powers are Hermitian, and hence also any analytic function (with real coefficients) of $\mathfrak{A}$ is Hermitian. Furthermore, a linear operator which is a function (with real coefficients) of several Hermitian and *commuting* linear operators is evidently itself Hermitian. However, if the linear operators of which it is a function do not commute, it cannot be Hermitian; for example, the linear operator $\mathfrak{A}\mathfrak{B} - \mathfrak{B}\mathfrak{A}$, if it is not zero, is not Hermitian, because otherwise $\mathfrak{D}$ would not be Hermitian as given by (50′).

10. Principal axes of an operator. Given a (Hermitian) linear operator $\mathfrak{A}$, we ask the following question: Are there vectors (in Hilbert space) which the operator $\mathfrak{A}$ changes in magnitude but not in direction? This amounts to a search for functions f such that

$$\mathfrak{A}f = Af, \tag{52}$$

where A is a multiplicative constant to be determined.

This problem, in the case where $\mathfrak{A}$ is the operator $\mathfrak{L}$ of (47) (with $A' = B$), consists in trying to find solutions of integrable square of the self-adjoint equation

$$Af'' + A'f' + (C - L)f = 0,$$

where L plays the role of "parameter." As we have seen in §2 of Part II, there exist an infinite number of independent solutions (*eigenfunctions*) $f = y_n$, to each of which there corresponds a value L_n of L (*eigenvalue*), some of the L_n possibly being equal to one another.

In general, we shall call *eigenvalues* of the linear operator $\mathfrak{A}$ the numbers A_n, and *eigenfunctions* the functions u_n such that

$$\mathfrak{A}u_n = A_n u_n, \tag{53}$$

and we shall always suppose that the u_n are normalized. Hence the latter represent as many unit vectors $\mathbf{u}_n$, which we may also suppose to be orthogonal to each other,[10] so that we shall have

---

[10] The proof of this fact for a general linear operator $\mathfrak{A}$ (provided it is Hermitian) is carried out in the same way as for the linear operator (47) in

$$\mathbf{u}_n \cdot \mathbf{u}_m = \delta_{nm}. \tag{54}$$

The directions of these vectors are called the *principal axes* of the linear operator $\mathfrak{A}$, and any vector lying along one of these axes is changed by the operator $\mathfrak{A}$ in magnitude only, not in direction.

The eigenvalues of a linear Hermitian operator are always real, as we shall show later on. They belong to a complete system of orthogonal functions and may be either continuous or discrete. We shall generally refer to the discrete case, it being understood that in all formulas having continuous eigenvalues we must replace the summations by appropriate integrals, and that we are to adopt as criterion for normalization and orthogonality the explanation in §10 of Part II.

Note on incomplete operators. In a problem in which several independent variables x, y, . . . occur, a linear operator is said to be *complete* if all the variables occur in its expression, and it is said to be *incomplete* if some of them are missing (for example, if y and the operations of differentiation with respect to y do not occur). We shall generally call x the variables which occur in the linear operator $\mathfrak{A}$, and y the variables which are missing. It is evident that by solving (53) we shall obtain for the u_n some functions $\chi_n(x)$ of x only, not of y. However, any one of these may be multiplied

---

§5 of Part II. If $\mathbf{u}_m$ and $\mathbf{u}_n$ belong to two different eigenvalues A_m and A_n, they are orthogonal. In fact, we have

$$\mathfrak{A}\mathbf{u}_n = A_n\mathbf{u}_n, \qquad \mathfrak{A}\mathbf{u}_m = A_m\mathbf{u}_m.$$

Multiplying the first of these through by $\mathbf{u}_m$ from the right and the second by $\mathbf{u}_n$ from the left (scalar product) and subtracting, we obtain

$$(\mathfrak{A}\mathbf{u}_n) \cdot \mathbf{u}_m - \mathbf{u}_n \cdot (\mathfrak{A}\mathbf{u}_m) = (A_n - A_m)\mathbf{u}_n \cdot \mathbf{u}_m.$$

But if $\mathfrak{A}$ is Hermitian, the left side is zero, and hence it follows (since $A_n \neq A_m$) that $\mathbf{u}_n \cdot \mathbf{u}_m = 0$, that is, that they are orthogonal.

If A_n is a multiple eigenvalue to which there belong the p independent eigenfunctions $u_n^{(1)}$, $u_n^{(2)}$, . . . $u_n^{(p)}$, the general eigenfunction belonging to this eigenvalue is of the form

$$u_n = \sum_{j=1}^{p} c_n{}^j u_n{}^j, \tag{55}$$

the coefficients c being arbitrary. Of course for the u_n we may substitute p of their linear combinations, which are orthogonal to each other (this is proved as in §6). There is still a large degree of arbitrariness in the choice of the combinations.

by an arbitrary function $f_n(y)$ without ceasing to satisfy (53). Because of this fact, $f_n(y)\chi_n(x)$ may also be considered as an eigenfunction $\varphi_n(x, y)$ of $\mathfrak{A}$, whose general expression will therefore be

$$\varphi_n(x, y) = f_n(y)\chi_n(x). \tag{56}$$

Hence, considering the Hilbert space of the functions of x and y, we can say that in this space the principal axes of the incomplete linear operator $\mathfrak{A}$ are not determined (even if possible degeneracies are ignored) but may be chosen with a large degree of arbitrariness by selecting the functions $f_n(y)$ at will. In general it will be convenient to assume the $f_n(y)$ to be orthonormal in y; thereupon (the χ_n being assumed to be normalized in x) the functions $\varphi_n(x, y) = f_n\chi_n$ will constitute an orthonormal system of functions (but not a complete system).

This indeterminacy in the eigenfunctions of an incomplete linear operator may in a way be likened to a degeneracy, with the eigenvalue A_n of $\mathfrak{A}$ considered as *multiple of infinite order.* In fact, let $u^j(y)$ be any complete system of orthogonal axes in the function space of y. The $f_n(y)$ mentioned above may be expanded in the form $f_n(y) = \sum f_n^j u^j$, and hence (56) may be written

$$\varphi_n(x, y) = \sum_{j=1}^{\infty} f_n^j u^j \chi_n = \sum_{j=1}^{\infty} f_n^j \varphi_n^j(x, y), \tag{57}$$

where we put $\varphi_n^j(x, y) = u^j(y)\chi_n(x)$. This formula is analogous to (55). It expresses the most general eigenfunction of $\mathfrak{A}$ belonging to the eigenvalue A_n, as a linear combination (with arbitrary coefficients f_n^j) of the fundamental eigenfunctions φ_n^j.

For applications the following theorem is of importance: if φ_n is an eigenfunction of $\mathfrak{A}$ belonging to the eigenvalue A_n, it is also an eigenfunction of $F(\mathfrak{A})$ belonging to the eigenvalue $F(A_n)$, where F stands for any analytic function (see §4).

Proof. By hypothesis, we have

$$\mathfrak{A}\varphi_n = A_n\varphi_n.$$

Operating upon both sides with the linear operator $\mathfrak{A}$, we obtain

$$\mathfrak{A}^2\varphi_n = A_n\mathfrak{A}\varphi_n = A_n^2\varphi_n.$$

Repeating this process, we recognize that the following expression is valid for any power of $\mathfrak{A}$:

$$\mathfrak{A}^r\varphi_n = A_n^r\varphi_n. \tag{58}$$

Assume now a function F defined by the series

$$F(\alpha) = \sum_r a_r\alpha^r,$$

and denote by $\mathfrak{F}$ the linear operator $F(\mathfrak{A})$. Then

$$\mathfrak{F}\varphi_n = \sum_r a_r\mathfrak{A}^r\varphi_n,$$

and because of (58),

$$\mathfrak{F}\varphi_n = \sum_r a_rA_n^r \cdot \varphi_n = F(A_n)\varphi_n, \tag{59}$$

which proves the theorem.

It is also evident that if the function F can be inverted (that is, if we may write $\mathfrak{A} = G(\mathfrak{F})$, where G stands for an analytic function), the converse of this theorem also holds true. In this case, then, we may say that the linear operators $\mathfrak{A}$ and $\mathfrak{F} = F(\mathfrak{A})$ have the same eigenvalues and the same eigenfunctions.

In the case of continuous eigenvalues it is important to observe that if $\varphi_{A'}$ is an eigenfunction of $\mathfrak{A}$ belonging to the eigenvalue A' and is *normalized* (according to the criterion of §10, Part II), it is also an eigenfunction of $F(\mathfrak{A})$ belonging to the eigenvalue $F(A')$ but is not normalized. To normalize it, it must be divided by $\sqrt{dF/dA'}$, as may easily be verified.

This theorem suggests an important generalization of the concept of function of a linear operator, which up to now has been limited to analytic functions. Let $F(\alpha)$ stand for a function of the variable α (possibly not developable in a series), and let $\mathfrak{A}$ be a linear operator with eigenvalues A_n and eigenfunctions φ_n. We shall define the linear operator $F(\mathfrak{A})$ as the operator which, when applied to a φ_n, transforms it into $F(A_n)\varphi_n$, in accordance with (59), and hence when applied to any f possessing the expansion

$$f = \sum_n a_n\varphi_n$$

transforms it into

$$\mathfrak{A}f = \sum_n a_nF(A_n)\varphi_n. \tag{60}$$

The only condition to which the function $F(\alpha)$ is subject is that it be uniquely defined for all values A_n of α.

11. Fundamental theorem on commuting operators. The necessary and sufficient condition for two linear operators $\mathfrak{A}$ and $\mathfrak{B}$ to have a complete system of eigenfunctions (and hence of principal axes) in common, is that they commute.

First we shall show that this condition is necessary. Let us suppose that there exists a complete system of orthogonal eigenfunctions $\varphi_1, \varphi_2, \ldots$ which are at the same time eigenfunctions of $\mathfrak{A}$ and $\mathfrak{B}$, so that we may write (numbering the subscripts of the eigenvalues suitably)

$$\mathfrak{A}\varphi_i = A_i\varphi_i, \qquad \mathfrak{B}\varphi_i = B_i\varphi_i,$$

where some of the A_i may be equal to each other (and likewise for the B_i), so as to include cases of degeneracy.

Operating upon the first equation with the linear operator $\mathfrak{B}$ and upon the second one with $\mathfrak{A}$, we obtain, respectively,

$$\mathfrak{B}\mathfrak{A}\varphi_i = A_i\mathfrak{B}\varphi_i = A_iB_i\varphi_i,$$
$$\mathfrak{A}\mathfrak{B}\varphi_i = B_i\mathfrak{A}\varphi_i = B_iA_i\varphi_i,$$

and hence

$$\mathfrak{A}\mathfrak{B}\varphi_i = \mathfrak{B}\mathfrak{A}\varphi_i.$$

Since the φ_i form a complete set, any function f may be expanded in a series of the φ_i, and hence for any f the following expression holds:

$$\mathfrak{A}\mathfrak{B}f = \mathfrak{B}\mathfrak{A}f,$$

which means that

$$\mathfrak{A}\mathfrak{B} = \mathfrak{B}\mathfrak{A}. \tag{61}$$

Let us now prove that the condition is sufficient. Let us suppose that (61) holds and let us call φ_i a complete set of eigenfunctions of the operator $\mathfrak{A}$, such that

$$\mathfrak{A}\varphi_i = A_i\varphi_i,$$

and let ψ_j be a complete set of eigenfunctions of $\mathfrak{B}$ (all orthogonal to one another) for which

$$\mathfrak{B}\psi_j = B_j\psi_j.$$

Let us develop the φ_i in a series of the ψ_j:

$$\varphi_i = \sum_j c_{ij}\psi_j$$

and apply the operator $\mathfrak{A}F(\mathfrak{B})$ to both sides, where $F(\mathfrak{B})$ is any linear operator function of $\mathfrak{B}$. Then we have, making use of the theorem of §10,

$$\mathfrak{A}F(\mathfrak{B})\varphi_i = \sum_j c_{ij}F(B_j)\mathfrak{A}\psi_j. \tag{62}$$

On the other hand, since $\mathfrak{A}$ commutes with $\mathfrak{B}$, it also commutes with $F(\mathfrak{B})$, and hence

$$\begin{aligned}\mathfrak{A}F(\mathfrak{B})\varphi_i &= F(\mathfrak{B})\mathfrak{A}\varphi_i = F(\mathfrak{B})A_i\varphi_i \\ &= A_i\sum_j c_{ij}F(\mathfrak{B})\psi_j = A_i\sum_j c_{ij}F(B_j)\psi_j.\end{aligned}$$

Comparing this with (62), we obtain

$$\sum_j c_{ij}F(B_j)\mathfrak{A}\psi_j = A_i\sum_j c_{ij}F(B_j)\psi_j, \tag{63}$$

and since F is an arbitrary function, the coefficients $F(B_j)$ must be considered to be entirely arbitrary numbers. It follows that equation (63) cannot hold, unless for each term in which $c_{ij} \neq 0$,

$$\mathfrak{A}\psi_j = A_i\psi_j. \qquad (j = 1, 2, \ldots) \tag{63'}$$

We now observe, however, that for a given j at least one of the coefficients c_{ij} must be $\neq 0$. Otherwise, in fact, the ψ_i in question would be orthogonal to all the φ_i (since $c_{ij} = \varphi_i \cdot \psi_j$), which is impossible because the φ_i constitute a complete set. It follows further that for any ψ_j there exists at least one index i for which (63′) holds, or that each ψ_j is an eigenfunction of $\mathfrak{A}$ (belonging to the eigenvalue A_i). Hence we have proved that the ψ_j make up a complete set of eigenfunctions which are common to $\mathfrak{A}$ and $\mathfrak{B}$.

Note. The preceding proof also holds for the cases in which multiple eigenvalues are present. Let us focus our attention particularly upon the meaning of this theorem in this case. If, for instance, A_i is a multiple eigenvalue of $\mathfrak{A}$, of order p, one may select, in an infinite number of ways, p orthogonal and independent eigenfunctions belonging to it (whose vectors form a plane manifold V in p dimensions). In general, however, for only one of these choices do we obtain eigenfunctions which are common to both operators. (It is apparent that to the unique eigenvalue A_i there correspond p eigenvalues of $\mathfrak{B}$; that is, there are p principal axes of $\mathfrak{B}$ within the subspace V.) This statement also applies in the case of infinite multiplicity, that is, when one of the operators, such as $\mathfrak{A}$, is incomplete.

In this case the theorem has the following implication. Let there be, for instance, two variables x and y, of which only x occurs in $\mathfrak{A}$. We already know that if we call $\chi_i(x)$ the eigenfunctions of $\mathfrak{A}$ in the function space of x, it is possible to consider, as eigenfunctions of $\mathfrak{A}$ in the function space of x and y, all the functions of the form $\varphi_i(x, y) = f(y)\chi_i(x)$, f being an arbitrary function of y alone. However, only some of these eigenfunctions are common to $\mathfrak{A}$ and $\mathfrak{B}$; specifically, $f(y)$ must be chosen from among a discrete (or possibly continuous) aggregate of functions $f_k(y)$, such that $\psi_{ik} = f_k(x)\chi_i(y)$, where we call ψ_{ik} the eigenfunctions of $\mathfrak{B}$ (we specify these eigenfunctions explicitly by two indices, because they depend on the two variables x and y). To each χ_i there correspond an infinite number of eigenfunctions of $\mathfrak{B}$.

12. Diagonal matrices. Let us take the principal axes of $\mathfrak{A}$ (with versors $\mathbf{u}_n$) for our coordinate axes in Hilbert space, and let us look for the form the matrix takes in representing the linear operator $\mathfrak{A}$ in reference to these axes. We shall designate this matrix by $\{\mathfrak{A}\}'$ and shall continue to denote by $\{\mathfrak{A}\}$ the matrix which represents $\mathfrak{A}$ with respect to the general axes y_n.

The general element A'_{mn} of the matrix $\{\mathfrak{A}\}'$ will be, according to (23),

$$A'_{mn} = \mathfrak{A}\mathbf{u}_n \cdot \mathbf{u}_m,$$

and, because of (53) and (54),

$$A'_{mn} = A_n\mathbf{u}_n \cdot \mathbf{u}_m = A_n\delta_{nm}. \tag{64}$$

In other words, the matrix $\{\mathfrak{A}\}'$ will be

$$\{\mathfrak{A}\}' = \begin{Vmatrix} A_1 & 0 & 0 & \cdots \\ 0 & A_2 & 0 & \cdots \\ 0 & 0 & A_3 & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{Vmatrix}.$$

A matrix such as this one, in which all the elements are zero except for those along the principal diagonal (*diagonal elements*), is called a *diagonal matrix*. Hence we can say: *A linear operator is represented, with respect to its principal axes, by a diagonal matrix; the diagonal elements of this matrix are the eigenvalues of the operator.*

It is apparent that if the operator is Hermitian, the matrix will also be Hermitian, which means, for a diagonal matrix, that its elements are real. Hence *the eigenvalues of a linear Hermitian operator are always real.*

Let there be given a linear operator $\mathfrak{A}$ by means of the matrix $\{\mathfrak{A}\}$ representing it in the general axes y_n. We want to find the (diagonal) matrix $\{\mathfrak{A}\}'$ which represents $\mathfrak{A}$ with respect to its principal axes. This operation is sometimes called *reduction of the matrix* $\{\mathfrak{A}\}$ *to the diagonal form* or *diagonalization of the matrix* $\{\mathfrak{A}\}$.

We recall from §8 that $\{\mathfrak{A}\}'$ is obtained from $\{\mathfrak{A}\}$ by formula (44). We are to find the transformation matrix $\{\mathfrak{S}\}$. For this purpose we note that (44) may also be written

$$\{\mathfrak{S}\}\{\mathfrak{A}'\} = \{\mathfrak{A}\}\{\mathfrak{S}\},$$

or else, by equating the general term (m, k) on both sides,

$$\sum_r S_{mr}A'_{rk} = \sum_r A_{mr}S_{rk};$$

but since $A'_{rk} = A_k\delta_{rk}$, the first summation reduces to the one term in which $r = k$, so that the equation becomes

$$\sum_r A_{mr}S_{rk} - A_kS_{mk} = 0.$$

Let us keep k fixed and assign to m the successive values $1, 2, \ldots$; we then obtain the equations

$$\left.\begin{array}{l} (A_{11} - A_k)S_{1k} + A_{12}S_{2k} + A_{13}S_{3k} + \cdots = 0 \\ A_{21}S_{1k} + (A_{22} - A_k)S_{2k} + A_{23}S_{3k} + \cdots = 0 \\ A_{31}S_{1k} + A_{32}S_{2k} + (A_{33} - A_k)S_{3k} + \cdots = 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \end{array}\right\}. \tag{65}$$

These constitute a system of infinitely many linear homogeneous equations in the infinite number of unknowns $S_{1k}, S_{2k}, S_{3k}, \ldots$

In order to understand the nature of this problem, let us first consider the case which occurs in ordinary three-dimensional space. Here the system (65) reduces to a system of three linear homogeneous equations in the three unknowns S_{1k}, S_{2k}, S_{3k} (we could write three of these systems, corresponding to $k = 1, k = 2, k = 3$), and possesses nonzero solutions only if

$$\begin{vmatrix} A_{11} - A_k & A_{12} & A_{13} \\ A_{21} & A_{22} - A_k & A_{23} \\ A_{31} & A_{32} & A_{33} - A_k \end{vmatrix} = 0. \tag{66}$$

This equation in A_k, called *secular equation* (which is the same for all three systems), possesses, as is known from algebra, three roots which are always real, A_1, A_2, A_3. Once found, they may be substituted into the three systems, and we can easily find the S_{mk} to within a factor which is determined by the conditions (34). This operation is seen to be identical with the classical procedure for the determination of the principal axes of a quadric surface (whose coefficients for the second degree terms are the A_{ij}). The S_{mk} furnish the direction cosines of these axes, and the lengths of the three semiaxes are given by $\pm 1/\sqrt{A_k}$.

Hence it may be seen that the problem of reducing a matrix to its diagonal form (or of finding the eigenvalues and eigenfunctions of a linear operator $\mathfrak{A}$) is the generalization, to the case of infinitely many dimensions, of the problem of the reduction of a quadric surface to its principal axes. Generalizing what is known for a system of n linear homogeneous equations with n unknowns to the case where n is infinite, we are led to the conclusion that the system (65) has nonzero solutions only if A_k has certain definite values (forming an infinite and generally discrete sequence) which are just the desired eigenvalues.

13. Matrices with continuous indices. Throughout the preceding sections we have assumed that the eigenfunctions which define the coordinate axes in Hilbert space form a discrete sequence. But, as we have said in §10 of Part II, when the region within which the differential equation is to be solved is infinite, eigenvalues forming continuous spectra may occur (besides possible discrete eigenvalues). Thus we are led to consider cases in which the indices occurring in the previous formulas vary continuously rather than taking on only integral values. Extension of the previous considerations to such cases contains some delicate points, for a rigorous treatment of which we refer to special works.[11] We shall restrict ourselves here to general remarks of an intuitive nature.

First we observe that if a_n is any sequence of quantities corresponding to the integral values of the index n, when the latter index becomes a continuous variable λ we must consider a_λ as standing for an ordinary function of λ, which may thus also be written $a(\lambda)$. Hence, as the eigenfunctions y_n are replaced by y_λ (λ is a continuous

---

[11] See, for instance, in addition to the works of Dirac, the paper by E. H. Kennard in *Zeits. f. Physik* **44**, 326 (1927).

index), every vector **f** in Hilbert space is represented, with respect to the axes $\mathbf{y}_\lambda$, by the following function of λ instead of by the components f_n:

$$\mathrm{f}_\lambda = \mathbf{f} \cdot \mathbf{y}_\lambda$$

(component of **f** along the $\mathbf{y}_\lambda$-axis). Similarly, all formulas of the theory of vectors in Hilbert space will be modified by substituting integrals for the summations with a single index. For example, (10) becomes

$$\mathbf{f} \cdot \mathbf{g} = \int_a^b f_\lambda g_\lambda^* \, d\lambda.$$

We shall now see what becomes of the matrices in this case. An expression with two subscripts A_{mn}, when m and n become two continuous variables μ, λ in the interval (a, b), will become an ordinary function of two variables, which may be written in the usual manner $A(\mu, \lambda)$ instead of $A_{\mu\lambda}$. Therefore the concept of "matrix with continuous indices" (or "continuous matrix") is equivalent to the concept of function of two variables. If we want to retain the same physical picture of the matrix, we must interpret μ and λ as Cartesian coordinates in a plane, taking the axes as in Fig. 46. Then to each point of the crosshatched square we may make correspond a value of $A(\mu, \lambda)$. We may say that the elements of such a matrix "continuously fill" the crosshatched square (which may eventually be extended to infinity, becoming a "quadrant" if $b = \infty$).

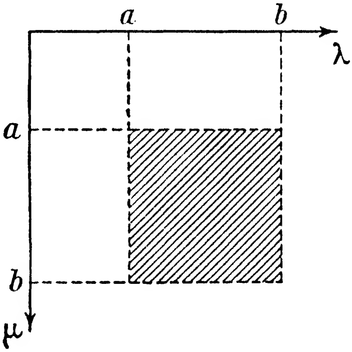

Fig. 46

All the definitions that have already been given may be extended to these continuous matrices. For instance, the product of the two matrices $A(\mu, \lambda)$ and $B(\rho, \sigma)$ is the matrix

$$P(\mu, \sigma) = \int_a^b A(\mu, \nu)B(\nu, \sigma) \, d\nu. \tag{67}$$

A continuous matrix is said to be *Hermitian* if

$$A^*(\mu, \lambda) = A(\lambda, \mu).$$

A Hermitian operator $\mathfrak{A}$ is represented, with respect to the axes specified by the eigenfunctions $\mathbf{y}_\lambda$, by just such a matrix, whose general element is, in analogy to (23) and (23′),

$$A(\mu, \lambda) = (\mathfrak{A}\mathbf{y}_\lambda) \cdot \mathbf{y}_\mu = \int y_\mu{}^* \mathfrak{A} y_\lambda \, dx.$$

By means of this continuous matrix, the components of the vector $\mathbf{F} = \mathfrak{A}\mathbf{f}$ are obtained from the components of $\mathbf{f}$ by the formula, analogous to (22),

$$F_\mu = \int_a^b A(\mu, \lambda) f_\lambda \, d\lambda.$$

In order to pass from one system of axes $\mathbf{y}_\lambda$ to another system $\mathbf{y}_\lambda{}'$, we must introduce a (continuous) *transformation matrix* defined by [see (33)]

$$S(\mu, \lambda) = \mathbf{y}_\lambda{}' \cdot \mathbf{y}_\mu.$$

By means of this matrix, we may pass from the components f of the vector $\mathbf{f}$ to the components f' of the same vector with respect to the new axes, by means of a formula analogous to (35):

$$f_\lambda{}' = \int_a^b S^*(\mu, \lambda) f_\mu \, d\mu. \tag{68}$$

The inverse transformation is given by [see (35″)]

$$f_\lambda = \int_a^b S(\mu, \lambda) f_\mu{}' \, d\mu. \tag{68′}$$

In this change of axes, by means of (44) the matrix $A(k, j)$ representing an operator $\mathfrak{A}$ with respect to the axes $\mathbf{y}_\lambda$ changes over into the matrix $A'(\mu, \lambda)$ representing the same operator in the new axes. Formula (44), when written explicitly, becomes the following, analogous to (43):

$$A'(\mu, \lambda) = \int_a^b \int_a^b S^*(\mu, k) A(k, j) S(j, \lambda) \, dk \, dj. \tag{69}$$

In conclusion, we shall mention the case in which the equation defining the y has discrete eigenvalues as well as a continuous spectrum between a and b. Here the previous formulas are modified in such a way that to each integral between a and b there must be added a summation over the discrete eigenvalues. This change introduces some complication into the writing (which may actually be avoided by the use of Stieltjes integrals; see No. 14 of the Bibliography, page 122) but adds no conceptual difficulties.

14. The Dirac δ-function. The extension to the case of continuous indices of the unit matrix introduced in §5 (that is, the generalization of the symbol δ_{mn}) leads us to introduce a symbol $\delta(x)$ which is rather convenient in calculations, and is called the *Dirac δ-function.* It represents a function which has the following properties:

$$\delta(x) = 0 \qquad \text{for} \qquad x \neq 0, \tag{70}$$

while at the point 0, δ is infinite in such a way that

$$\int_a^b \delta(x)\,dx = 1, \tag{71}$$

where a and b are any two limits ($a < b$), containing 0 between them. Strictly speaking, no function possessing these properties exists, and hence $\delta(x)$ is called an *improper function.* We may approximate it as closely as we like, however, by means of analytic functions.[12] For example, the Gaussian function $h/\sqrt{\pi}\, e^{-x^2/h^2}$ for $h \to \infty$ approaches the shape of $\delta(x)$. Hence we may imagine the curve representing $\delta(x)$ as an infinitely narrow and infinitely high bell-shaped curve, enclosing an area equal to unity.

We shall generally make use of the function $\delta(x - x_0)$, which exhibits at the point $x = x_0$ the same singularity $\delta(x)$ has at $x = 0$. It has the fundamental property that, if $f(x)$ is any function (provided it is bounded over the interval considered and continuous at x_0), then

$$\int_a^b f(x)\,\delta(x - x_0)\,dx = \begin{cases} 0 \text{ if the interval } (a, b) \text{ does not} \\ \quad \text{contain } x_0, \\ f(x_0) \text{ if the interval } (a, b) \text{ does} \\ \quad \text{contain } x_0. \end{cases} \tag{72}$$

In fact, if the interval does not contain x_0, the function δ is zero over the entire interval, and hence the integral is zero. If the interval (a, b) contains x_0, the integral may be limited to a segment

---

[12] The use of the improper function may be avoided by replacing it by the Stieltjes integral concept, as was done systematically by Neumann (see No. 13 of the Bibliography). Its use, however, contributes to making the formulas simpler and more expressive, and may be considered as the abbreviated indication of a passage to the limit.

$x_0 - \epsilon$, $x_0 + \epsilon$ (ϵ being as small as desired). Then, calling f_1 a value lying between the lower and upper limits of f in the interval $(x_0 - \epsilon, x_0 + \epsilon)$, we have

$$\int_a^b f(x)\,\delta(x - x_0)\,dx = \int_{x_0-\epsilon}^{x_0+\epsilon} f(x)\,\delta(x - x_0)\,dx$$
$$= f_1 \int_{x_0-\epsilon}^{x_0+\epsilon} \delta(x - x_0)\,dx = f_1;$$

and since this must hold for any ϵ, we must have $f_1 = f(x_0)$.

A property of δ which we shall often use is expressed by the formula

$$(x - x_0)\delta(x - x_0) = 0, \tag{73}$$

which is to be justified by saying that for all values of x for which the first factor does not vanish, the second factor does vanish. More precisely, we say that (73) is to be understood in the sense that the integral of the left side with respect to x, extended over any interval (as small as desired), is zero.

The introduction of the improper function δ makes possible a formal consideration of the axes of Hilbert space, which we have called "continuous" in §1, as principal axes of a linear operator, that is, as a particular case of the $\mathbf{y}_n$-axes considered up to now. In fact, we consider the linear operator $\mathfrak{x} = x$ and look for its principal axes, writing the equation

$$x\varphi_{x'}(x) = x'\varphi_{x'}(x), \tag{74}$$

or

$$(x - x')\varphi_{x'} = 0, \tag{74'}$$

where we have indicated the eigenvalue by x' (since we are dealing with continuous eigenvalues, as we shall see). Now (74′) is satisfied by taking any x' and

$$\varphi_{x'}(x) = \delta(x - x'), \tag{75}$$

as is shown by (73). In addition, the eigenfunctions are orthonormal, since (see §10 of Part II), upon calling $\Delta_1 x'$, $\Delta_2 x'$ two infinitesimal intervals, we have, as we easily see by using (72),

$$\int_{-\infty}^{\infty} dx \int_{\Delta_1 x'} \delta(x - x')\,dx' \int_{\Delta_2 x'} \delta(x - x')\,dx'$$
$$= \begin{cases} 0 \text{ if } \Delta_1 x', \Delta_2 x' \text{ have no point in common} \\ \Delta x \text{ if } \Delta_1 x', \Delta_2 x' \text{ have a segment } \Delta x \text{ in common.} \end{cases} \tag{76}$$

Therefore: the eigenvalues of the operator x are all the real numbers x', and to each of these there corresponds an axis specified by the function (75). These axes are actually the "continuous axes" introduced in §1. In fact, the projection of any function $f(x)$ upon the axis corresponding to the eigenvalue x' is

$$f_{x'} = \mathbf{f} \cdot \boldsymbol{\delta}(x - x') = \int f(x)\delta(x - x')\,dx = f(x'). \tag{77}$$

From the formal standpoint, it is apparent that the symbol δ_{nm} which is often used ($\delta_{nm} = 0$ if $n \neq m$, $\delta_{nm} = 1$ if $n = m$) has its analogue in the function $\delta(\lambda - \mu)$. For instance, the condition for orthogonality and for normalization of the eigenfunctions, expressed by (46) and (46′) of Chapter 5 may be written symbolically as

$$\int y_{\lambda'} y^*_{\lambda''}\,dx = \delta(\lambda'' - \lambda'). \tag{78}$$

Indeed, if (78) is satisfied, we have (calling $\Delta\lambda'$, $\Delta\lambda''$ two infinitesimal sections of the continuous eigenvalue spectrum)

$$\begin{aligned}\int_{-\infty}^{\infty} dx \int_{\Delta\lambda'} y_{\lambda'}\,d\lambda' \int_{\Delta\lambda''} y^*_{\lambda''}\,d\lambda'' &= \int_{\Delta\lambda'} d\lambda' \int_{\Delta\lambda''} d\lambda'' \int_{-\infty}^{\infty} y_{\lambda'} y^*_{\lambda''}\,dx \\ &= \int_{\Delta\lambda'} d\lambda' \int_{\Delta\lambda''} d\lambda''\,\delta(\lambda'' - \lambda').\end{aligned}$$

The integral with respect to λ'' is 1 if λ' lies within the small interval $\Delta\lambda''$; otherwise it is zero. Therefore, if we call $\Delta\lambda$ the segment (possibly zero) common to the two intervals $\Delta\lambda'$, $\Delta\lambda''$, the last expression reduces to

$$\int_{\Delta\lambda} d\lambda',$$

that is, to $\Delta\lambda$. Thus again we find the conditions for orthogonality and normalization introduced in §10 of Part II.

CHAPTER 11

General Theory of Quantum Mechanics

15. The concept of "observable." We shall call *observation* (or *measurement*) a series of physical operations whose result is expressible by means of a number (we shall include in the operations of measurement the eventual mathematical operations to be performed upon the direct experimental results). An observation will be defined when these operations are described and when the instant in which they are performed is specified; for example, one of the experimental arrangements of §23, Part II, with a specification of the time t at which the photograph is to be taken, constitutes an *observation.*

In ordinary mechanics, an observation serves to find the value possessed by a certain "mechanical quantity" at a certain instant, that is, a coordinate, a momentum, or any function $g(q, p)$ of the coordinates and momenta, such as, for instance, the angular momentum or the energy. This value, however, is thought of as existing even if we do not perform the measurement which is designed to find it. In other words, g is thought of as a *variable* which takes on different numerical values in succession during the motion of the system; these values may be calculated a priori, even without observation, by means of the laws of mechanics, if, for example, the initial coordinates and velocities of the system are known.

In quantum mechanics, as has been noted in connection with the coordinates of a particle, we must adopt a profoundly different point of view. The numerical value of a physical quantity has no meaning whatsoever until an observation upon it is performed, and the result of that observation may in general only be predicted probabilistically. In other words, it is maintained that the result of the observation may be any one of certain numbers[1] $G_1, G_2, \ldots$

---

[1] For simplicity of writing, the case considered here is the one with discrete eigenvalues, but the G may also constitute a continuous system, for instance in the measurement of a coordinate.

with respective probabilities[2] $P_1, P_2, \ldots$; the object of quantum mechanics is just the determination of these possible results and their respective probabilities. Examples of this method were seen in connection with the observation of a coordinate or a momentum (§25 of Part II), or of the energy (§29 of Part II).

In quantum mechanics, then, entities such as the coordinates, the components of momentum, the energy, and so on (at a given instant), are considered from a point of view different from that of ordinary mechanics; that is, they are entities defined only by the type of "observation" which corresponds to them. In order to make this condition evident, Dirac has proposed the name *observables* for these quantities. (We recall that in the concept of *observation*, and hence also in the concept of *observable*, there is implicit also the instant at which the measurement is performed; for example, the abscissa of a particle at the time $t = 7$ sec is an observable,[3] and the abscissa of the same particle at the time $t = 8$ sec is *another* observable).

In ordinary mechanics, the concept of observable is synonymous with that of "numerical value of a variable." In quantum mechanics, however, an observable does not in general have a numerical value, but there correspond to it an infinity of numerical values (continuous or discrete) $G_1, G_2, \ldots$ with respective probabilities $P_1, P_2, \ldots$. Only in special cases may it happen that these probabilities are all zero except one, for example, P_1 (which is therefore $= 1$), and only then can we say that the observable G *has* the value G_1. For instance, this case is realized, for the energy, when the system is in one of the states which we have called "stationary" or "states with definite energy" (§27 of Part II), that is, when ψ is an eigenfunction of the Schrödinger equation. However, when ψ is a linear combination of eigenfunctions (§29 of Part II), the energy has no definite numerical value. A direct observation designed to determine it may yield for result any one of the eigenvalues $E_1, E_2, \ldots$ with the respective probabilities $c_1c_1^*, c_2c_2^* \ldots$.

---

[2] The concept of probability must be understood in the manner explained in the footnote on page 145.

[3] Strictly speaking, we should specify the measuring arrangement; for example, the pinhole camera described in §23 of Part II (first method) defines an observable which we may call abscissa; it is a generally accepted postulate that the second and third methods define the same observable.

One more example: the coordinate x of a particle (at a given time) is in general an observable which has no numerical value, since by observing it we may find all the values from $-\infty$ to $+\infty$, with a probability density $P(x) = \psi\psi^*$. The curve representing $P(x)$ may have any form whatever, some examples of which were seen in §36 of Part II. In particular, $P(x)$ may be zero everywhere, except for a definite x, and only in this case does the observable x have a definite value.

16. Compatibility of observables. Let us suppose that the measurement of an observable A has yielded a certain result A', and let us measure, immediately following A, another observable B, obtaining B', for example. By this measurement we put the system in a new state in which B has the definite value B', but A in general no longer will have a definite value (that is, if immediately after B we returned to an observation of A, we could no longer be certain of finding A'). There are, however, certain cases in which this does not hold true—those in which we may observe B right after A, without A ceasing to have the value resulting from the previous measurement. The two measurements are then said to be *compatible*. In this case, just after the observation of A has yielded the result A' and the observation of B has given B', the system finds itself in a state such that both A and B have definite values, namely, A' and B' respectively.

As may be seen, the preceding definition applies only to *simultaneous* observations. The question therefore arises: does it make any difference whether we perform the observation of A first and the observation of B immediately afterward, or the other way around? We shall see in what follows that the laws of quantum mechanics require the former choice, which we shall assume for now without further remarks.

Two *observables* are said to be *compatible* if their observations are always compatible, no matter what their results. This is one of the most important concepts of quantum mechanics. An example of two compatible observables is given by two coordinates of a particle (for example, x and y) at the same instant.[4] On the other hand, a coordinate x and the conjugate momentum p_x are not compatible, as was seen in §22 of Part II.

---

[4] At least if the processes of measurement by which these observables are defined satisfy certain conditions (see §24).

Hence we shall say that several observables A, B, C, . . . are compatible if each of them is compatible with all the others.

17. Observables which are functions of other observables. Given an observable X and a function $f(\lambda)$, the observable $G = f(X)$ is defined by the following set of operations. First we perform the observation X, and upon the result X_r we perform the mathematical operations indicated by the function f. By definition, the number $G_r = f(X_r)$ is the result of the measurement of the observable G. We may also say that G is measured with the same instrument which measures X except for the substitution (for the scale calibrated in terms of the numbers X_r) of another scale on which the numbers $f(X_r)$ are marked in the same positions.

It is clear that if we suppose f to be single-valued and if the G_r are discrete, their probabilities will be equal to those of the corresponding quantities X_r. If, however, X has a continuous spectrum of values X', with probability densities $P(X')$, the values G' of the observable G form a continuous spectrum with probability density

$$Q(G') = \frac{P(X')}{df/dX'}.$$

It follows readily from the definition that an observable X is compatible with any $f(X)$.

We now proceed to the definition of a function of several observables X, Y, Z, . . . (referring to the same instant). If X, Y, Z are compatible with each other, the procedure for the definition of $F(X, Y, Z, \ldots)$ is the immediate extension of the procedure for the case of a single variable: a simultaneous measurement of X, Y, Z, . . . , yielding numerical results X_r, Y_r, Z_r, . . . , followed by the mathematical operations expressed by the function $F(X_r, Y_r, Z_r, \ldots)$.

It is then quite evident that the observable F defined in this manner is compatible with each of the observables X, Y, Z, ...

However, in the case that X, Y, Z, . . . are not compatible with each other, this procedure is evidently no longer applicable. Nevertheless, given a function of several variables $F(x, y, z, \ldots)$, we can define (at least under rather broad conditions) an observable $F(X, Y, Z, \ldots)$ in the following manner.

Let us first of all consider the case of the sum of several observables X, Y, Z, . . . which are not compatible. Given an observable G (defined by any procedure) we shall give the following criterion for deciding whether or not G is equal to the sum $X + Y + Z + \cdots$. We examine a large number of copies of the system under consideration (all equally constituted but possibly in different states); then we select a very large number of these at random and perform a measurement of X upon each system of this group. Let $\bar{X}$ be the average of the results obtained. We then select another very large group at random, in which we measure Y and compute $\bar{Y}$, and so forth; finally we take one last, large group in which we determine G and calculate $\bar{G}$. If it now happens (for any initial state of the ensemble of systems) that

$$\bar{G} = \bar{X} + \bar{Y} + \bar{Z} + \cdots,$$

we shall say that G is the sum of $X, Y, Z, \ldots$, and we shall write[5]

$$G = X + Y + Z + \cdots.$$

It is apparent that this definition, when $X, Y, Z, \ldots$ are compatible with each other, does not conflict with the definition given above for the function of several compatible observables. In fact, this definition evidently preserves all the ordinary properties of this operation.

From the definition of the sum we may pass on to the definition of "symmetrized product," that is, of $\frac{1}{2}(XY + YX)$. Indeed, supposing that there exists an observable G such that $G = X + Y$, the formal properties of algebra are preserved if we can write $G^2 = X^2 + Y^2 + XY + YX$. Since X^2, Y^2, G^2 represent observables which have already been defined, this reasoning leads us to introduce the following definition for the symmetrized product:

$$\tfrac{1}{2}(XY + YX) = \tfrac{1}{2}(G^2 - X^2 - Y^2);$$

---

[5] This definition does not permit us to construct the procedure of measurement for G if we know those of $X, Y, Z, \ldots$. Hence there remains the question whether there exists in every case an observable G which has the indicated property. However, in the cases which occur in practice, this condition may actually be verified, and we shall postulate that it is always verifiable. For example, the kinetic energy of an electron is defined by $X = (1/2m)(p_x^2 + p_y^2 + p_z^2)$ and hence is obtainable from a measurement of momentum; the potential energy is defined by $Y = eV(x, y, z)$ and can be obtained from a measurement of position. They are incompatible observables, but the total energy E (definable in reference to spectroscopic terms) has the property $\bar{E} = \bar{X} + \bar{Y}$, and therefore we may write $E = X + Y$.

that is, we shall say that an observable F is equal to $\frac{1}{2}(XY + YX)$ if it is equal (in the sense specified above) to $\frac{1}{2}(G^2 - X^2 - Y^2)$. If the observables X and Y are compatible, their symmetrized product is identified with the product XY or YX. However, if they are incompatible, no meaning can be given to the expressions XY or YX, but only to $\frac{1}{2}(XY + YX)$. By an analogous procedure, the symmetrized products for any desired number of factors may be defined.

Now given a function of several variables $F(x, y, z, \ldots)$ developable in a power series, we may always write each term of the series in symmetrized form and therefore may attach a meaning to the form which is obtained by replacing x, y, z, and so on, by as many observables (even if not compatible) $X, Y, Z, \ldots$. In this way the meaning of $F(X, Y, Z, \ldots)$ is defined.

18. Maximum observations. State of a system. In classical mechanics, when the positions and velocities of all the points of a system at a given instant have been assigned, its "state" at the instant considered has been *completely* defined, meaning that any further condition (for example, the assignment of a given value to the energy) would be either incompatible with the preceding ones or else automatically satisfied.

We shall now see how this notion may be carried over into quantum mechanics. If, for example, we assign values to the coordinates of a system of particles at a given instant, we have already reached a complete description of the state of the system, in the sense explained above. In fact, any further condition would be either automatically satisfied or incompatible with the others. (Thus, if we were also to assign the value of a component of velocity, this condition would be incompatible with the preceding ones by virtue of the uncertainty principle.) Hence we shall say that a measurement of all the coordinates of the particles constitutes a "complete set of observations," or a "maximum observation," because it furnishes, so to speak, a maximum amount of information obtainable from measurements performed on the system. Instead of the coordinates, we could measure all the components of momentum, obtaining in this way another complete set of observations. In general, we shall say that a set of observations is *complete*, or that it constitutes a *maximum observation*, if there exists no other observation which is independent and compatible with the set considered.

The results of a maximum observation define completely the *state* of the system (which is supposed to be isolated). Thus the quantum mechanical notion of "state" is precisely described. It is to be noted that the state defined in this manner does not refer to a particular instant but rather to the entire time interval in which the system remains isolated. During that time, of course, the system evolves according to a well-determined process characteristic of the state in question. Thus, for example, if the three coordinates of a particle have been measured at the time $t = 0$, the state of the particle for all later times (until it interacts with other systems, possibly with a radiation field) remains the same, and is defined by the fact that for $t = 0$ the coordinates had those given values. (Naturally, this does not mean that the coordinates retain these values; rather, they suddenly cease to have definite values.) To each state, there corresponds a certain function ψ which characterizes it, as we shall see later when we generalize what has been said for a single particle.

If one of the observations serving to define the state is a measurement of energy, we are dealing with one of the states which we have called (§27 of Part II) *stationary states,*[6] or *quantum states,* to which there correspond the eigenfunctions of the Schrödinger equation.

All that has been said so far applies to isolated systems. In many cases, however, the notion of state may be extended also to systems under external influences, where the state of the system at a given instant t is defined as the state in which the system would be left if the external action were to cease at the time t. It is to be understood that in general the state of a system which is not isolated will vary in time.

Returning now to the case of an isolated system, we note that in order to perform any observation upon it, we must necessarily make it interact with another system (that is, with the measuring apparatus) and hence we will interrupt its isolation at least momentarily. Therefore, *an observation performed upon a system generally alters its state.* However, there are cases in which there is no alteration; for instance, if the system is in a stationary state we

---

[6] This designation, the reason for which we shall see in §24, is not to suggest that these states are the only ones which do not vary in time. For instance, if we superimpose two stationary states by taking for ψ a linear combination of two Schrödinger eigenfunctions (see §29 of Part II), we get a *nonstationary* state which is nevertheless invariant in time.

may measure its energy (again finding the initial value, of course) without altering its state.

Let us compare the quantum-mechanical notions of "maximum observation" and "state" with the corresponding classical concepts. We observe first of all that in classical mechanics, when we, know the state of the system at time $t = 0$ (in the sense specified at the beginning of this section), we can calculate the value of any mechanical quantity relative to the system. Hence, for example a knowledge of the coordinates and velocities at time $t = 0$ is equivalent to a knowledge of all mechanical variables at the time $t = 0$ (and also at any other instant). In quantum mechanics, on the other hand, when a maximum observation has been performed, generally we may not calculate the result of a further observation (even if it is performed at the same instant) but may only assign the possible results and their relative probabilities (by a method whose general form will be explained in this chapter). Therefore a complete knowledge of the "state" of the system does not imply any knowledge of the aggregate of all quantities relative to the system (or else we may not attribute to this aggregate any physical significance). In general there are only a few observables which have definite values, whereas only probabilistic indications may be given concerning the others.

Another essential difference between classical and quantum mechanics resides in the following. If in classical mechanics we were to substitute for the group of values of the coordinates and velocities another group of as many of their independent functions, we should obtain a representation of the state which would be equivalent to the other in all respects. In quantum mechanics, however, there are an infinite number of nonequivalent ways of giving a complete description of the state of the system, depending on which complete set of observations is selected. For example, to assign the values of the coordinates or to assign the values of the momenta are two equally complete ways of defining the state of the system. They are not equivalent, however, since the observables having a definite value in the first case do not have a definite value in the second case, and vice versa. In this connection, see also §32.

19. Interpretation of the Schrödinger method in Hilbert space. We shall now briefly recapitulate the procedure of the wave mechanics of Schrödinger, stating it in the geometrical terminology of Hilbert space. In this manner the route to many important generalizations will be perceived.

We start out by noticing a formal analogy between the Schrödinger equation for stationary states, which we write in the form

$$\left[-\frac{h^2}{8\pi^2 m}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)+U\right]\psi_n = E_n\psi_n, \tag{79}$$

and the kinetic energy theorem of classical mechanics, which we write

$$\frac{1}{2m}(p_x^2+p_y^2+p_z^2)+U = E. \tag{80}$$

The analogy consists in the following: if in (80) we replace bodily the variables

$$\left.\begin{array}{c} p_x, \qquad\qquad p_y, \qquad\qquad p_z, \\ \text{respectively by the (Hermitian) } operators \\ \mathfrak{p}_x = \dfrac{h}{2\pi i}\dfrac{\partial}{\partial x}, \qquad \mathfrak{p}_y = \dfrac{h}{2\pi i}\dfrac{\partial}{\partial y}, \qquad \mathfrak{p}_z = \dfrac{h}{2\pi i}\dfrac{\partial}{\partial z} \end{array}\right\}, \tag{S}$$

the left side is changed into the operator which is applied to ψ_n in the left-hand member of (79).

Recalling from the footnote on page 234, that the expression for the energy as a function of the q and the p was generally indicated by $\mathcal{H}(q, p)$ and called the *Hamiltonian*, we shall say that the Schrödinger equation is obtained by transforming the Hamiltonian $\mathcal{H}$ into an operator $\mathfrak{H}$ (which we shall call *Hamiltonian operator*) by means of the substitution (S) and writing

$$\mathfrak{H}\psi_n = E_n\psi_n; \tag{81}$$

that is (see §10), we are trying to find the eigenvalues and eigenfunctions of this operator (which turns out to be Hermitian). Its eigenvalues E_n represent the possible values of the energy, and its eigenfunctions ψ_n, which may be interpreted as orthogonal unit vectors directed along the principal axes of the operator $\mathfrak{H}$, have the following physical interpretation. When the system is in the nth stationary state and we perform an observation of the coordinates, the probability density of finding the values x, y, z is given by $|\psi_n(x, y, z)|^2$; in geometrical terms, by the square of the modulus of the projection of the vector[7] ψ_n upon the axis (of the system of continuous axes; see §2) specified by the values x, y, z. For brevity we shall say that the "probability amplitude" for the values x, y, z is equal to this projection.

When the system is not in a stationary state, that is, when ψ has the general form $\psi = \sum_n c_n\psi_n$, or when the vector ψ is not directed along one of the principal axes of the operator $\mathfrak{H}$, the same meaning still holds for the components of that vector along the continuous axes. Furthermore, if we perform a measurement of energy, the probability of finding the value E_n is $|c_n|^2$ (see §29 of Part II); that is, the amplitude of this probability is c_n, which is the projection of the vector ψ upon the nth principal axis of the operator $\mathfrak{H}$.

Hence the vector ψ determines the probability of the results of any measurement of the coordinates or of the energy (and also, as we shall see later, of any other observation carried out on the system). The energy is generally a function of time, and its evolution in time is governed by the time-dependent Schrödinger equation (see §30 of Part II), which may now be written in the form

$$\mathfrak{H}\psi = -\frac{h}{2\pi i}\frac{\partial\psi}{\partial t}. \tag{82}$$

This equation may be considered as obtained from the classical equation for the conservation of energy ($\mathfrak{H} = E$), where we perform, in addition to the substitution (S), the substitution of

$$E \quad \text{by} \quad -\frac{h}{2\pi i}\frac{\partial\psi}{\partial t} \tag{S'}$$

and operate upon ψ with the operators thus obtained.

The vector ψ, considered as a function of time, characterizes the *state* of the system. It will henceforth be called *state vector.*

20. Extension to a system of N distinct particles. The preceding section's formulation of the Schrödinger problem for a single

---

[7] For simplicity, we shall use the same letter to indicate a function ψ and the corresponding vector in Hilbert space (rather than using a boldface symbol for the latter, as in the preceding chapter).

particle suggests, by an obvious generalization, a method for treating the more general problem of any number of *distinct*[8] particles.

Let N be the number of particles; their $3N = f$ coordinates will sometimes be indicated by $x_1, y_1, z_1, x_2, y_2, z_2, \ldots, z_N$, and sometimes, when it is convenient, by $q_1, q_2, \ldots q_f$. We must now introduce (generalizing the criterion for the procedure followed in §25 of Part II) the probability that a simultaneous observation of *all* the particles at the time t yields values for the coordinates lying between $q_1, q_2, \ldots q_f$ and $q_1 + dq_1, q_2 + dq_2, \ldots q_f + dq_f$, respectively. This probability will be denoted by

$$P(q_1, q_2, \ldots q_f, t)dq_1\,dq_2 \ldots dq_f.$$

If, on the other hand, we carry out an observation on the kth particle only, without considering the others, the probability of finding it in the element of volume defined by $x_k, y_k, z_k, x_k + dx_k, y_k + dy_k, z_k + dz_k$ will be

$$P_k(x_k, y_k, z_k, t)dx_k\,dy_k\,dz_k,$$

where P_k will be obtained from P by integrating the latter with respect to all coordinates except x_k, y_k, z_k, and over all values which these coordinates may assume.

It is evident that the N functions P_k are defined once P is known, but in general P is not determined from a knowledge of the P_k. This statement may be understood, apart from analytical means, by the following intuitive consideration. Let there be a box with two equal compartments A and B, into which two small balls a and b are thrown at random, the balls being *independent of each other*. If we observe only one of these, the probability of finding it in either A or B is ½. If we observe both of them, the following four cases may occur:

a in A, b in B:	probability	¼
a in A, b " A:	"	¼
a in B, b " B:	"	¼
a in B, b " A:	"	¼.

Let us now consider that the two balls are so large that only one can enter each compartment. Then there is still a probability of ½ for each to be found in either A or B; but for both of them taken together, the probabilities of the preceding table become, respectively, ½, 0, 0, ½. This

---

[8] Here "distinct" is to mean that we suppose each particle to have its own individuality; that is, we may distinguish it from the others. If we were dealing with identical particles (electrons, for example), other arguments would have to be made; these will be postponed until Chapter 15.

result shows that the knowledge of the relative probabilities of the separate observations of each particle is not sufficient to determine the relative probabilities of the observations of both of them together. In fact, it is also necessary to know how the position of one particle influences the probability of the position of the other. If this influence is zero (as in the first case), we say that the two balls are "statistically independent."

Returning to the case of N elementary particles, we shall say that the latter are statistically independent if $P(q_1, q_2, \dots q_f, t)$ has the following form:[9]

$$P = P_1(x_1, y_1, z_1, t)P_2(x_2, y_2, z_2, t) \dots P_N(x_N, y_N, z_N, t). \tag{83}$$

This is the only case in which P is determined from a knowledge of the P_k. It is to be noted that the particles cannot be statistically independent even if no forces exist between them. This is a very important point, to which we shall return later on.

In analogy to what was done for a single particle, we introduce a (complex) function

$$\psi(q_1, q_2, \dots q_f, t)$$

such that

$$P(q_1, q_2, \dots q_f, t) = |\psi|^2,$$

and impose upon ψ the condition to satisfy an equation which is to be the generalization of the Schrödinger equation for a single particle. This generalization will be based on the analogy mentioned in §19. For this purpose we start from the classical expression for the Hamiltonian of a system of N particles in Cartesian coordinates. Let us suppose that forces[10] act upon them which are derivable from a potential $U(q_1, q_2, \dots q_f)$, and let us indicate by $p_x^{(k)}, p_y^{(k)}, p_z^{(k)}$ the momenta conjugate to x_k, y_k, z_k (or by p_j the momentum conjugate to q_j). The mass of the kth particle will be denoted by $m^{(k)}$, and its charge by $e^{(k)}$. The Hamiltonian of the system may then be written

$$\mathcal{H} = \sum_{k=1}^{N} \frac{1}{2m^{(k)}} (p_x^{(k)2} + p_y^{(k)2} + p_z^{(k)2}) + U(x_1, y_1, \dots z_N),$$

---

[9] It may be verified immediately that by integrating P with respect to all variables, except x_k, y_k, z_k, over all their ranges of variation, we obtain $P_k(x_k, y_k, z_k)$.

[10] We neglect the magnetic interactions between the particles of the system, which are intimately connected with the relativistic corrections which will be introduced in Chapter 14.

or else, in a more convenient notation,

$$\mathcal{H} = \sum_{j=1}^{3N} \frac{1}{2m_j} p_j^2 + U(q_1, q_2 \ldots q_{3N}), \tag{84}$$

where $m_1 = m_2 = m_3$ stand for $m^{(1)}$, $m_4 = m_5 = m_6$ stand for $m^{(2)}$, and so on.

Let us transform this expression into an operator, as was done in the case of a single particle, by replacing

$$p \quad \text{by} \quad \mathfrak{p}_j = \frac{h}{2\pi i} \frac{\partial}{\partial q_j}, \tag{85}$$

whereupon we have

$$\mathfrak{H} = -\frac{h^2}{8\pi^2} \sum_{j=1}^{f} \frac{1}{m_j} \frac{\partial^2}{\partial q_j^2} + U(q_1, q_2, \ldots q_f)$$

or else, indicating by

$$\Delta_k = \frac{\partial^2}{\partial x_k^2} + \frac{\partial^2}{\partial y_k^2} + \frac{\partial^2}{\partial z_k^2}$$

the Laplacian operator for the kth particle,

$$\mathfrak{H} = -\sum_{k=1}^{N} \frac{h^2}{8\pi^2 m_k} \Delta_k + U(q_1, q_2, \ldots q_f). \tag{86}$$

Let us now assume that the ψ of the system satisfies the following equation, which is a generalization of the time-dependent Schrödinger equation [see (136) of Part II]:

$$\mathfrak{H}\psi = -\frac{h}{2\pi i} \frac{\partial \psi}{\partial t}, \tag{87}$$

or

$$\sum_{k=1}^{N} \frac{1}{m_k} \Delta_k \psi - \frac{8\pi^2}{h} U\psi = -\frac{4\pi i}{h} \frac{\partial \psi}{\partial t}. \tag{87'}$$

The stationary states or states with definite energy will be those for which ψ is an eigenfunction ψ_n of the operator $\mathfrak{H}$, that is, such that[11]

$$\mathfrak{H}\psi_n = E_n \psi_n. \tag{88}$$

---

[11] Here, of course, the index n stands for a group of f indices.

The time dependence of these ψ_n is obtained by comparing (88) with (87), which yields

$$\frac{\partial \psi_n}{\partial t} = -\frac{2\pi i}{h} E_n \psi_n$$

and hence
$$\psi_n = u_n(q_j)\, e^{-\frac{2\pi i}{h} E_n t}, \tag{89}$$

where $u_n(q_j)$ satisfies the equation

$$\mathfrak{H} u_n = E_n u_n, \tag{88'}$$

as in the case of a single particle.

The most general ψ may of course be developed in a series of the ψ_n of (89); that is, any state of the system may be considered as a superposition of stationary states.

It is interesting to verify that equation (87), which was adopted by induction, is satisfied, in particular, in the case of N *dynamically independent* particles—those which do not exert forces on each other but are subject only to external forces. In this case, U evidently breaks up into a sum of N functions U_k, each of which contains only the coordinates of one particle, and hence the operator (86) may also be split into a sum of N operators $\mathfrak{H}_k$, each operating on the coordinates of only one particle:

$$\mathfrak{H}_k = -\frac{h^2}{8\pi^2 m_k} \Delta_k + U_k.$$

Hence (87) becomes

$$\sum_{k=1}^{N} \mathfrak{H}_k \psi = -\frac{h}{2\pi i} \frac{\partial \psi}{\partial t}. \tag{87''}$$

Now we can see immediately that this equation *can* be satisfied by taking

$$\psi = \psi^{(1)} \psi^{(2)} \cdots \psi^{(N)}, \tag{90}$$

where $\psi^{(k)}$ is a function of x_k, y_k, z_k, t only and satisfies the equation

$$\mathfrak{H}_k \psi^{(k)} = -\frac{h}{2\pi i} \frac{\partial \psi^{(k)}}{\partial t},$$

which is the ordinary Schrödinger equation referring to the kth particle. For the ψ of the system, then, we may take the product of the ψ of the separate particles. In this case, (83) is evidently satisfied; that is, the particles are statistically independent. If each of the particles is in a stationary state (with index n_k), that is, if

$$\mathfrak{H}\psi_{nk}^{(k)} = E_{nk}^{(k)}\,\psi_{nk}^{(k)},$$

we may verify immediately that ψ, which is the product of all the $\psi_{nk}^{(k)}$, satisfies the equation

$$\sum_k \mathfrak{H}_k\psi = E_{n_1n_2\ldots n_N}\psi,$$

with

$$E_{n_1n_2\ldots n_N} = \sum_{k=1}^{N} E_{nk}^{(k)};$$

this amounts to saying that the system is in a stationary state, and that its energy is the sum of the energies of the individual particles.

Hence the splitting of the Hamiltonian into a sum of N terms, each of which depends on only the coordinates of a single particle, entails the possibility of splitting ψ into the product of N factors, each corresponding to one particle, and of breaking up every eigenvalue into the sum of N terms, each representing the energy of one particle.

But it is to be noted that (87″) (established under the single hypothesis that the particles are dynamically independent) admits, in addition to solutions of the type (90), any linear combination of such solutions. These solutions represent the cases in which the particles are not statistically independent, although they are dynamically independent.

It is easily recognized that if the particles are statistically independent at a given instant t_0, they are also statistically independent at any other instant (provided, of course, that no forces act between them). In fact, because of (87″), ψ is uniquely determined by its values for $t = t_0$ in all space; hence if solution (90) holds for $t = t_0$, it holds also for any other time t.

It is to be noted that whereas in the case of a single particle ψ represents fictitious waves in ordinary space, usually in the case of N particles we cannot retain this picture, since ψ contains $3N$ coordinates (in addition to t). Therefore ψ may be interpreted only by means of waves in a $3N$-dimensional space.

21. The two-body problem. As an application, let us consider the problem which arises in the treatment of hydrogenlike systems when we wish to take into account the fact that the nucleus is not fixed, as was assumed in §48 of Part II. This problem, which has been treated in §58 of Part II from the viewpoint of the Bohr-Sommerfeld theory, corresponds to the well-known two-body problem in classical mechanics.

Let x_1, y_1, z_1 be the coordinates of the nucleus and x_2, y_2, z_2 the coordinates of the electron (with respect to any fixed axes), and let p_{1x}, p_{1y}, p_{1z}, p_{2x}, p_{2y}, p_{2z} be the respective momenta conjugate to these coordinates. Let M be the mass of the nucleus and m the mass of the electron. The classical Hamiltonian is

$$\mathcal{H} = \frac{1}{2M}(p_{1x}^2 + p_{1y}^2 + p_{1z}^2) + \frac{1}{2m}(p_{2x}^2 + p_{2y}^2 + p_{2z}^2) + U, \quad (91)$$

where U is the potential energy, which, since it depends only on the relative position, will be a function of $(x_2 - x_1)$, $(y_2 - y_1)$, $(z_2 - z_1)$.

Therefore $\psi(x_1, y_1, z_1, x_2, y_2, z_2, t)$ must satisfy the equation

$$\mathfrak{H}\psi = -\frac{h}{2\pi i}\frac{\partial\psi}{\partial t}, \quad (92)$$

with

$$\mathfrak{H} = -\frac{h^2}{8\pi^2 M}\left(\frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial y_1^2} + \frac{\partial^2}{\partial z_1^2}\right) - \frac{h^2}{8\pi^2 m}\left(\frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial y_2^2} + \frac{\partial^2}{\partial z_2^2}\right) + U(x_2 - x_1, y_2 - y_1, z_2 - z_1). \quad (93)$$

Let us express this operator, not in terms of the variables x_1, y_1, z_1, x_2, y_2, z_2, but by using the three center-of-mass coordinates

$$\xi = \frac{Mx_1 + mx_2}{M + m}, \qquad \eta = \frac{My_1 + my_2}{M + m}, \qquad \zeta = \frac{Mz_1 + mz_2}{M + m},$$

and the three coordinates of the electron with respect to the nucleus

$$x = x_2 - x_1, \qquad y = y_2 - y_1, \qquad z = z_2 - z_1.$$

Evidently we have

$$\frac{\partial}{\partial x_1} = \frac{\partial\xi}{\partial x_1}\frac{\partial}{\partial\xi} + \frac{\partial x}{\partial x_1}\frac{\partial}{\partial x} = \frac{M}{M + m}\frac{\partial}{\partial\xi} - \frac{\partial}{\partial x},$$

and hence

$$\frac{\partial^2}{\partial x_1^2} = \left(\frac{\partial}{\partial x_1}\right)^2 = \left(\frac{M}{M+m}\right)^2 \frac{\partial^2}{\partial \xi^2} + \frac{\partial^2}{\partial x^2} - \frac{2M}{M+m}\frac{\partial^2}{\partial \xi\, \partial x}.$$

Similarly, we find

$$\frac{\partial^2}{\partial x_2^2} = \left(\frac{\partial}{\partial x_2}\right)^2 = \left(\frac{m}{M+m}\right)^2 \frac{\partial^2}{\partial \xi^2} + \frac{\partial^2}{\partial x^2} + \frac{2m}{M+m}\frac{\partial^2}{\partial \xi\, \partial x}.$$

By substituting into (93), we see that the terms with mixed derivatives cancel, and similarly for the other coordinates, so that we are left with

$$\mathfrak{H} = -\frac{h^2}{8\pi^2(M+m)}\left(\frac{\partial^2}{\partial \xi^2} + \frac{\partial^2}{\partial \eta^2} + \frac{\partial^2}{\partial \zeta^2}\right) - \frac{h^2}{8\pi^2}\left(\frac{1}{m} + \frac{1}{M}\right)\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right) + U(x, y, z).$$

Since U does not contain ξ, η, ζ, the operator may be split into a sum of two parts, one of which, $\mathfrak{H}_0$, contains only ξ, η, ζ, the other of which, $\mathfrak{H}'$, contains only x, y, z:

$$\mathfrak{H}_0 = -\frac{h^2}{8\pi^2(M+m)}\left(\frac{\partial^2}{\partial \xi^2} + \frac{\partial^2}{\partial \eta^2} + \frac{\partial^2}{\partial \zeta^2}\right),$$

$$\mathfrak{H}' = -\frac{h^2}{8\pi^2 m'}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right) + U(x, y, z),$$

where we have put

$$\frac{1}{m'} = \frac{1}{m} + \frac{1}{M};$$

that is,
$$m' = \frac{mM}{M+m} = \frac{m}{1 + (m/M)}.$$

(m' is called the *reduced mass.*) Correspondingly, ψ may be split (see §20) into the product

$$\psi = \chi(\xi, \eta, \zeta)\Psi(x, y, z),$$

where the two functions χ and Ψ satisfy the equations

$$\mathfrak{H}_0\chi = -\frac{h}{2\pi i}\frac{\partial \chi}{\partial t}, \tag{94}$$

$$\mathfrak{H}'\Psi = -\frac{h}{2\pi i}\frac{\partial \Psi}{\partial t}. \tag{95}$$

As may be seen, the operator $\mathfrak{H}'$ is identical with the operator which occurs in the problem of the motion of an electron about a supposedly *fixed* nucleus, except for the substitution, for the mass m, of the (slightly smaller) mass m'. Hence Ψ will be identical with the ψ of the theory developed in §46 of Part II, provided that we replace the mass m by the reduced mass m'. (This is the same modification which must be made in classical mechanics or in the Bohr-Sommerfeld theory in order to account for the motion of the nucleus; see §58 of Part II.)

As far as the operator $\mathfrak{H}_0$ is concerned, it is the one which arises in the motion of a particle of mass $M + m$ not subject to forces. χ is therefore an eigenfunction of the free point particle (see §44 of Part II). This means that the motion of the center of mass may be treated, in wave mechanics as well, as the motion of a mass point of mass equal to the total mass of the system.[12]

As for the energy levels, they turn out to be (see the preceding section) the sum of an eigenvalue of (94) and an eigenvalue of (95). This result signifies that the energy of the atom in a stationary state may be divided into an energy of translation and an internal energy, just as in ordinary mechanics. For the latter energy (which alone is of interest in spectroscopy), then, the eigenvalues found in §48 of Part II must hold, with the slight correction introduced by replacing the mass m by the reduced mass m'.

22. Fundamental principle of quantum mechanics. Correspondence between observables and operators. In the procedure outlined in §§19 and 20, the observable "energy" occupies a privileged position. We may try to generalize these considerations by assigning to some other general observable G the role held by the energy up until now. The generalization will be made in analogy to the previous case, except that we are going to verify its validity by comparing with experiment the consequences deduced from that case.

For this purpose we shall assume the *postulate* that to each observable G we may make correspond a linear Hermitian operator $\mathfrak{G}$ which has the following properties: *its eigenvalues G_r represent*[13]

---

[12] This result makes it legitimate to apply the Schrödinger equation to the over-all motion of a complex system such as an atom or a molecule. The experiments on the diffraction of H, H_2, and He, confirm this result (see §29 of Part I).

[13] These eigenvalues always turn out real, since the operator is Hermitian.

the possible results of a measurement of G, *and its principal axes, specified by the eigenfunctions* φ_r *of the equation*

$$\mathfrak{G}\varphi_r = G_r\varphi_r, \tag{96}$$

have the property that the projection of the state vector ψ *upon the* rth *principal axis yields (supposing that* G_r *is not a multiple eigenvalue) the probability amplitude of obtaining the result* G_r *in a measurement of* G. This is equivalent to saying that if we call this component ψ_{G_r}, that is, if we set

$$\psi_{G_r} = \psi \cdot \varphi_r = \int \psi\varphi_r^* \, dS$$

(where the φ_r are assumed to be normalized), the probability of the value G_r is $|\psi_{G_r}|^2$. (Note that since the vector ψ is of unit magnitude, we have $\sum_r |\psi_{G_r}|^2 = 1$, as it must, in order that the above-mentioned significance of probability may be attributed to the $|\psi_{G_r}|^2$.) However, if the operator has a continuous eigenvalue spectrum, we shall indicate a general one of these continuous eigenvalues by G' rather than by G_r, and the corresponding eigenfunction by $\varphi_{G'}$ (normalized according to the criterion of §10 of Part II). Then $|\varphi_{G'}|^2 \, dG'$ will be the probability that an observation of G will give a result lying between G' and $G' + dG'$. In general, for the sake of convenience of writing, we shall refer here to the case of discrete eigenvalues, it being understood that obvious modifications are to be made in the formulas for the case of continuous eigenvalues.

If, in particular, the state vector lies along one of the principal axes of $\mathfrak{G}$ (that is, if $\psi = \varphi_r$), the system is in a state such that a measurement of G yields the value G_r with certainty. This happens for the energy, when the state is a *stationary state.*

Note on cases with degeneracy. If G_r is an eigenvalue of multiplicity p which has p corresponding orthogonal eigenfunctions $\varphi_r^1, \varphi_r^2, \ldots \varphi_r^p$, where ψ_r^i $(i = 1, 2, \ldots p)$ are called the projections of ψ upon these axes, we must take

$$P_r = \sum_{i=1}^{p} |\psi_r^i|^2 \tag{97}$$

for the probability of the value G_r. (We note that this equation is invariant with respect to any orthogonal transformation of the φ_r^i.) We may justify this step by considering the degenerate operator $\mathfrak{G}$ as a limit of a non-degenerate operator $\mathfrak{H}$, putting, for instance, $\mathfrak{H} = \mathfrak{G} + \epsilon\mathfrak{F}$ (where $\mathfrak{F}$ is chosen so as not to make $\mathfrak{H}$ degenerate, and where ϵ is a quantity which will be made to approach zero). Now to the eigenvalue G_r of $\mathfrak{G}$ there correspond p eigenvalues H_r^i $(i = 1, 2, \ldots p)$ of $\mathfrak{H}$ which coincide with G_r when ϵ goes to zero, and to these eigenvalues there correspond as many eigenfunctions $\bar{\varphi}_r^i$ and projections $\bar{\psi}_r^i$ of ψ. The probability of the value H_r^i for the observable H is $|\bar{\psi}_r^i|^2$, and hence the total probability for the p values $H_r^1, H_r^2, \ldots H_r^p$, which merge into G_r for $\epsilon \to 0$, is $\sum |\bar{\psi}_r^i|^2$. If we let ϵ go to zero, this expression has (97) for a limit.

An analogous case occurs if the operator $\mathfrak{G}$ is incomplete. For example, let x and y be the variables of the problem, and suppose that the operator $\mathfrak{G}$ contains only x. We know then (see §10) that to the eigenvalue G_r there correspond the infinite number of eigenfunctions $\varphi_r^j = u^j(y)\chi_r(x)$ (where the functions $u^j(y)$ represent any complete, orthonormal set of eigenfunctions in the function space of y). Therefore, according to (97), for the probability of the value G_r we shall assume[14]

$$P_r = \sum_{j=1}^{\infty} |\psi_r^j|^2, \tag{97'}$$

where $\psi_r^j = \psi \cdot \varphi_r^j$. This expression is independent of the choice of the u^j. Taking for these functions the functions $\delta(y - y')$, where y' takes the place of the index j (see §14), we have

$$\begin{aligned} \psi_r^{y'} &= \psi \cdot \delta(y - y')\chi_r(x) \\ &= \int\int \psi(x, y)\delta(y - y')\chi_r^*(x)\, dx\, dy \\ &= \int \psi(x, y')\chi_r^*(x)\, dx. \end{aligned}$$

Hence (97′) is transformed into the following integral (where we have written y instead of y'):

$$P_r = \int |\psi_r^y|^2\, dy. \tag{98}$$

From this we get the following *rule:* To obtain the probability P_r, calculate the eigenfunctions $\chi_r(x)$ of the operator $\mathfrak{G}$ in the function space of the functions of x alone, and expand the function $\psi(x, y)$ in a series of these eigenfunctions, considering it to be a function of x only. Of course the coefficients will be functions of y (and may therefore be denoted by ψ_r^y). The square of the modulus of the rth coefficient, *integrated with respect to* y, gives P_r.

Evidently, if $\mathfrak{G}$ had continuous eigenvalues G', in (98) we should have to substitute the continuous variable G' for the discrete index r, and we

---

[14] This assumption could be justified directly by a limiting procedure entirely analogous to that used in the case of multiple eigenvalues of order p.

should have to interpret $P_{G'}$ as the probability density; thus the probability for a value lying between G' and $G' + dG'$ would be

$$dG'\int|\psi^{y}_{G'}|^2\,dy. \tag{99}$$

It remains to be seen how we may determine the operator $\mathfrak{G}$ corresponding to a given observable G. For this purpose, we shall proceed by way of successive generalizations.

Case (a): $G = x$. If the observable G is one of the coordinates of the particle, x for example, the operator $\mathfrak{G}$ corresponding to it is, as we shall now show, the multiplication by x; that is, $\mathfrak{x} = x$ (it is incomplete, except for the one-dimensional case). In fact, as we have seen in §14, such an operator has for its eigenvalues all the real numbers x', and for eigenfunctions $\delta(x - x')$. The projection of the state vector ψ upon this principal axis (calculated in the function space of x) is

$$\psi_{x'} = \int\psi(x, y, z, t)\,\delta(x - x')\,dx = \psi(x', y, z, t),$$

and hence the probability of finding for x a value lying between x' and $x' + dx'$ is, by reason of (99),

$$dx'\int\int|\psi(x', y, z, t)|^2\,dy\,dz,$$

in full accord with the meaning of $|\psi|^2$ as a probability density.

Case (b): $G = p_x$. If the observable G is a component of the momentum, p_x for example, we may verify that the corresponding operator is

$$\mathfrak{p}_x = \frac{h}{2\pi i}\frac{\partial}{\partial x}.$$

In fact, relation (96) may then be written, p'_x being a general eigenvalue, as

$$\frac{h}{2\pi i}\frac{\partial\varphi_{p_x'}}{\partial x} = p'_x\varphi_{p_x'}.$$

This equation has any value of p'_x for eigenvalue. By a simple integration, it gives

$$\varphi_{p_x'} = \frac{1}{\sqrt{h}}\,e^{(2\pi i/h)p_x'x}, \tag{100}$$

where the factor $1/\sqrt{h}$ has been determined from the normalization condition.

The projection of ψ upon the principal axis p'_x is

$$\psi_{p_x'}(y, z, t) = \frac{1}{\sqrt{h}} \int_{-\infty}^{\infty} \psi(x, y, z, t)\, e^{-(2\pi i/h)p_x'x}\, dx, \tag{101}$$

and the probability that the x-component of the momentum lies between p'_x and $p'_x + dp'_x$ is, by virtue of (99),

$$dp'_x \int\int |\psi_{p_x'}|^2\, dy\, dz. \tag{102}$$

Now let us calculate the same probability by the principle of superposition. If we decompose ψ into Fourier integrals (considering it as a function of x only, or considering y and z to be fixed) we obtain

$$\psi = \frac{1}{\sqrt{h}} \int \psi_{p_x'}\, e^{(2\pi i/h)p_x'x}\, dp'_x, \tag{103}$$

where $\psi_{p_x'}$ is simply expression (101). Hence $\psi_{p_x'}$ measures the amplitude of the monochromatic component, of wavelength h/p'_x, of the de Broglie waves of a particle of *given* y and z (one-dimensional case; see §36 of Part II). Hence, by the principle of superposition, $|\psi_{p_x'}|^2\, dp'_x$ is the probability that the particle of given y and z has an x-component of momentum lying between p'_x and $p'_x + dp'_x$. Leaving y and z entirely undetermined, we evidently obtain just the value given by (102) for the probability that p_x lies between p'_x and $p'_x + dp'_x$.

Case (c): *an observable defined as a function of the coordinates or of the momenta.* Let A be an observable to which a certain operator $\mathfrak{A}$ is known to correspond, with eigenvalues A_r and eigenfunctions φ_r. Let G be another observable defined as a function of A, that is, $G = F(A)$. The possible results of a measurement of G will be $F(A_r)$, and each of these will have the same probability as the value A_r of A. Now it is easily seen that this result may also be found by applying the procedure of page 331 and making the operator $\mathfrak{G} = F(\mathfrak{A})$ correspond to G. In fact, as was seen in §10, this operator has the same principal axes as $\mathfrak{A}$ (specified by the vectors φ_r), with eigenvalues $G_r = F(A_r)$. Hence the projection of the state vector upon the rth axis of $\mathfrak{G}$ is the same ψ_r which represents the projection of ψ upon the rth axis of A; and thus the probability of the value G_r is $|\psi_r|^2$, like the probability of the value A_r of A. The theorem may immediately be extended to a function

$F(A, B, \ldots)$ of several observables compatible with one another, and it may be proved in the same way.

Now, having already seen that the operators corresponding to the coordinates q_r are the q_r themselves and that the operators corresponding to the momenta p_r are $(h/2\pi i)(\partial/\partial q_r)$, we can conclude that to an observable defined as $F(q)$ there corresponds the operator $F(q)$, and that to an observable $F(p)$ there corresponds $F[(h/2\pi i)(\partial/\partial q)]$, that is, the operator obtained from the function by substituting for each p_r the expression $(h/2\pi i)(\partial/\partial q_r)$.

(d) *General case.* Now let G be any observable, defined directly (that is, by operations which do not imply a measurement of the q and the p). This is the case of the energy, for instance. No general procedure exists for finding the corresponding operator $\mathfrak{G}$. However, in almost all the cases which occur in practice, the following heuristic method is successful; it is based upon the postulate of a profound formal analogy between ordinary mechanics and quantum mechanics. Of course this method will be justified a posteriori by the success of its consequences, which have been verified in all cases to which the method has been applied.

Ordinary mechanics permits us to express the value of the quantity G at the time t as a function $G = F(q, p)$ of the Cartesian coordinates q_i and of the momenta p_i (and possibly of the time t, which we now consider to be fixed). Let us suppose that this function may be expanded in a power series, and let us write it (if there are terms of the type $q_r^m p_r^n$) in symmetrized form, as explained in §17. After we have done this, the natural generalization of what we saw above for the case of $F(q)$ or $F(p)$ leads us to accept the following postulate:[15] *to the observable G which in classical mechanics has the (eventually symmetrized) expression $F(q, p)$, there corresponds the operator $\mathfrak{G} = F[q, (h/2\pi i)(\partial/\partial q)]$.* The cases already examined under (a), (b), and (c) are evidently included in this rule.

Finally, there are some cases in which it is impossible to make an expression of classical mechanics correspond to the observable in question. In these cases, the operator $\mathfrak{G}$ representing the observa-

---

[15] In what is to follow, we shall need to apply this postulate only to functions of the form $F(q, p) = Q(q) + P(p) + \sum_r A_r(q)p_r$, where only the last part requires symmetrization.

ble is introduced by means of an *ad hoc* hypothesis, which is to be justified by its consequences. Such is the case, for instance, of the components of spin in the Pauli theory (see §45).

To sum up, the fundamental principle of quantum mechanics may be stated as follows. Once the operator $\mathfrak{G}$ corresponding to the observable G is determined, either by means of the rule cited above or by means of a special hypothesis, we write the equation

$$\mathfrak{G}\varphi = G'\varphi, \tag{104}$$

where G' is a parameter. The eigenvalues G_r of this equation furnish the possible results of a measurement of the observable G; and if we call φ_r the corresponding eigenfunctions, the probability of the value G_r in the state ψ is given by $|\psi \cdot \varphi_r|^2$. For cases of degeneracy or of continuous eigenvalues, see page 332.

It is immediately apparent that if we apply this general rule to the case where the observable G is the energy of a particle, or of a system of particles, subject to forces derivable from a potential [in which case the function $F(q, p)$ is the Hamiltonian $\mathcal{H}(q, p)$], we again find the procedure of §19 for deriving the Schrödinger equation for stationary states, or its generalization (88). Hence this equation now appears as a particular case of the general problem of the search for the possible values of an observable G, and of the corresponding probabilities. The particular case in question occupies a privileged position in quantum mechanics because of the fundamental importance of the observable "energy" and also because of its property of remaining constant.

23. Note on the use of curvilinear coordinates. The rule of the preceding section for finding the operator corresponding to a given observable G presupposes that the G is expressed in terms of the *Cartesian* coordinates q_r and of the conjugate momenta p_r. In many cases, however, it is more convenient to obtain the classical expression for G by means of some generalized coordinates, which we shall denote by Q_r, calling P_r the conjugate momenta. Let $F(Q, P)$ be that expression. In order to obtain the operator G, it is necessary to go to Cartesian coordinates, to perform the substitution (S), and then once more to transform the operator obtained by expressing it in terms of the Q. This process is not, in general, equivalent to the simple substitution of $(h/2\pi i)(\partial/\partial Q_r)$ for P_r in $F(Q, P)$. (An example of this procedure has been given in §21.)

Take, for instance, the case of a point particle in a plane and not subject to any forces. If we use the polar coordinates $Q_1 = r$, $Q_2 = \theta$, and the respective momenta $P_1 = m\dot{r}$, $P_2 = mr^2\dot{\theta}$, the expression for the Hamiltonian is

$$\mathcal{H} = \frac{1}{2m}\left(P_1^2 + \frac{P_2^2}{r^2}\right). \tag{105}$$

In Cartesian coordinates, on the other hand, $\mathcal{H} = (1/2m)(p_x^2 + p_y^2)$, and the corresponding operator is the familiar

$$\mathfrak{H} = -\frac{h^2}{8\pi^2 m}\Delta,$$

which, when expressed in polar coordinates, is written as follows:

$$\mathfrak{H} = -\frac{h^2}{8\pi^2 m}\left[\frac{1}{r}\frac{\partial}{\partial r}r\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial \theta^2}\right].$$

This expression is not obtained from (105) but from its algebraic equivalent:

$$\mathcal{H} = \frac{1}{2m}\left[\frac{1}{r}P_1 r P_1 + \frac{1}{r^2}P_2^2\right], \tag{105'}$$

by the substitution of $(h/2\pi i)(\partial/\partial Q_r)$ for P_r.

24. Some consequences of the fundamental principle of quantum mechanics. Let us suppose that the measurement of an observable G has furnished the result G_r. If we indicate by ψ^- the state vector immediately prior to the observation, it is obvious that immediately after the observation, the state of the system will be specified by a vector ψ^+, which is in general different from ψ^-; that is, the perturbation produced in the system by the observation will have caused a sudden change of the state vector. We recognize immediately that ψ^+ must lie along a principal axis of the operator $\mathfrak{G}$, an axis which corresponds to the eigenvalue G_r. In fact, the immediate repetition of the observation G would yield the result G_r with certainty, and hence the projections of ψ^+ upon the principal axes of $\mathfrak{G}$ corresponding to the eigenvalues different from G_r must be zero, or ψ^+ must be orthogonal to all the other axes. Now first let us suppose that to the eigenvalue G_r there corresponds a single axis (that is, that the operator $\mathfrak{G}$ is complete and the eigenvalue G_r simple). The vector ψ^+ will then be specified (it is sufficient to know its direction, since we already know that it is a unit vector); that is, the state of the system after the observation is completely determined by the result of the observation, without our having to know anything about the state prior to the observation.[16] Such observations, so to speak, *completely renew* the state of the system and permit it to be defined by a number. A typical example is the measurement of the energy of a nondegenerate system, in which case the states specified in this manner are what we have called "stationary states"[17] or "definite-energy states." We may now characterize them by the property that the state vector lies along one of the principal axes of the Hamiltonian operator $\mathfrak{H}$. On the other hand, the states which we have called "states with indefinite energy" and which we have characterized in §29 of Part II, taking for ψ a linear combination of eigenfunctions ψ_n, are represented by a vector which does not lie along any of the principal axes of $\mathfrak{H}$. However, another complete Hermitian operator $\mathfrak{G}$ may exist (in fact, an infinite number of them do exist),[18] one of whose principal axes coincides with the direction of ψ; or (assuming that for every Hermitian operator there is a corresponding observable) a real observable G (other than the energy) exists such that the state under consideration (with indefinite energy) may be thought of as a "state with well-defined G" and may be characterized by the value G_r of G.

---

[16] It is apparent that once the vector ψ^+ has been determined at the instant immediately following the observation, the further evolution of ψ in time is governed by the time-dependent Schrödinger equation, until a new observation perturbs the system.

[17] The reader will now immediately understand the reason for this name by noting that for one of these states the vector ψ has the form $\psi_n = u_n e^{-(2\pi i/h)E_n t}$, and hence maintains its direction and modulus unchanged in time (although its "phase" varies).

[18] In the one-dimensional case, for instance, if we let $\psi = \rho e^{i\theta}$ [with $\rho(x)$ and $\theta(x)$ real], it may readily be verified that this condition is satisfied by the operator

$$G(x, \mathfrak{p}) = \left(\mathfrak{p} - \frac{h}{2\pi}\theta'\right)^2 + \frac{h^2}{4\pi^2}\frac{\rho''}{\rho}$$

and that the ψ corresponds to the eigenvalue 0. Of course, any function of this G also satisfies the desired condition (see E. Fermi, *Nuovo Cimento* **VII, 10,** 361 (1930)

Let us now briefly consider the cases in which more than one principal axis corresponds to the eigenvalue G_r, either because G_r is a multiple eigenvalue or because $\mathfrak{G}$ is an incomplete operator. In these cases ψ^+ will not be determined by the result of the observation but will also depend on some details of the observation which are not reflected in its numerical result.[19] In order to establish these details precisely, we shall make the hypothesis that the observation is to be performed in such a way as to "perturb the system as little as possible," by which we understand the following. One expands ψ^- in a series of the eigenfunctions φ_r of $\mathfrak{G}$. Into this expansion there will enter, for each eigenvalue G_r, a certain one of the infinitely many eigenfunctions belonging to this eigenvalue. We shall say that the observation is carried out with minimum possible perturbation if ψ^+ is identical with this given eigenfunction. Since we shall always suppose the observations carried out under this condition, we may assert that in every case, ψ^- may be expanded in a series of all the possible ψ^+.

The geometrical interpretation of what has just been said is as follows. If the eigenvalue G_r is multiple of order p, there correspond to it an infinite number of principal axes which form a linear manifold in p dimensions (orthogonal to all the other principal axes). The vector ψ^+ must in every case lie within this manifold, which we shall call V. In what we have termed the case of minimum perturbation, ψ^+ is identical with the projection of ψ^- upon the manifold V.

The case of an incomplete operator may be made to fall within the preceding class by considering p to be infinite.

25. Criterion for the compatibility of observables. We shall now prove a theorem of the greatest importance: *The necessary and sufficient condition for two observables to be compatible is that their operators commute.*

Let A and B be two observables (relative to the same instant) which are compatible. Let us measure A, obtaining A_r, for instance; and then B, obtaining B_s. Then the system is left in a state in

---

[19] For instance, if we measure x of a particle without simultaneously determining y and z (or p_y and p_z), the result of the measurement is not enough to determine the succeeding state of the system, since some momentum is generally imparted along the y- and z-directions while x is being measured (for example, with a pinhole camera; see §23 of Part II). This is reflected in the fact that the operator x is incomplete. We may, however, arrange the experiment so as to be certain not to alter p_y and p_z, but only p_x, and the observation will then be carried out "with minimum possible perturbation."

which a subsequent (immediate) observation of A would yield the result A_r with certainty, and an observation of B would certainly give B_s. Hence the state vector ψ^+ after the double observation will lie along a principal axis common to the two operators $\mathfrak{A}$ and $\mathfrak{B}$. Since this statement must hold true no matter what the preceding state, the vector ψ^-, no matter what its form, must be developable in eigenfunctions common to $\mathfrak{A}$ and $\mathfrak{B}$, or these eigenfunctions must form a complete system. Therefore, by the theorem of §11, $\mathfrak{A}$ and $\mathfrak{B}$ commute.

Conversely, if $\mathfrak{A}$ and $\mathfrak{B}$ commute, they possess a complete system of principal axes in common (see §11). First let us suppose that neither $\mathfrak{A}$ nor $\mathfrak{B}$ is degenerate. Then any principal axis of $\mathfrak{A}$ is also a principal axis of $\mathfrak{B}$, to which we may assign the same index; hence, after we have performed the observation of A with the result A_r, the state vector will be lying along the principal axis corresponding to the value A_r of $\mathfrak{A}$ and B_r of $\mathfrak{B}$. Therefore by observing B we certainly obtain B_r, and the state vector is not altered by this new observation. The same holds true if we first observe B and then A. Consequently, the two observables are compatible. In fact, since to any value of the one there corresponds a definite value of the other, they are functions of each other.

Now let us suppose that $\mathfrak{B}$ is a degenerate (or incomplete) operator. Then the observation of A aligns the state vector in the direction of the A_r-axis (hence ψ becomes φ_r), which is again one of the infinite number of axes of $\mathfrak{B}$ belonging to a certain eigenvalue B_s which is multiple of order p (possibly ∞). These axes constitute a plane manifold V orthogonal to all other axes of $\mathfrak{B}$ belonging to other eigenvalues. Hence the subsequent observation of B must of necessity yield the result B_s and at most will rotate the state vector within V. However, since we suppose that the observation is carried out with minimum possible perturbation (in the sense explained above), the new state vector ψ^+ must be the orthogonal projection of φ_r upon V, which is evidently identical with φ_r itself. Hence ψ^+ will at the same time lie along a principal axis of $\mathfrak{A}$ and $\mathfrak{B}$. If we now were to observe B first (with result B_s) and then A (with result A_r), the first observation would bring the state vector into the manifold V; and since (see §11) within V there are p principal axes of $\mathfrak{A}$, whereas all the other principal axes of $\mathfrak{A}$ are perpendicular to V, the subsequent observation of A will necessarily align the state vector with one of the p-axes contained in V (and hence common to A and B). It is to be noted that to a value of B there may correspond p values of A.

Analogous reasoning is applied to the case in which both operators are degenerate or incomplete; here the relation between the results of the two measurements is even less stringent, or is lacking altogether.

In particular it follows from this theorem that the condition for the compatibility of two observations is symmetric, as was pointed out in §16.

26. Commutation relations. By means of the theorem just proved, it is possible to arrive again at the well-known fact that a Cartesian coordinate q_i ($i = 1, 2, 3$) and the corresponding component of momentum p_i are incompatible observables. Indeed, we have seen that their respective operators are

$$\mathfrak{q}_i = q_i, \qquad \mathfrak{p}_i = \frac{h}{2\pi i}\frac{\partial}{\partial q_i},$$

and evidently do not commute, since we have for any function f

$$\mathfrak{p}_i\mathfrak{q}_i f - \mathfrak{q}_i\mathfrak{p}_i f = \frac{h}{2\pi i}\frac{\partial}{\partial q_i}(q_i f) - \frac{h}{2\pi i}q_i\frac{\partial f}{\partial q_i} = \frac{h}{2\pi i}f,$$

or

$$\mathfrak{p}_i\mathfrak{q}_i - \mathfrak{q}_i\mathfrak{p}_i = \frac{h}{2\pi i}. \tag{106}$$

On the other hand, a $\mathfrak{q}_i$ and a $\mathfrak{p}_j$ ($j \neq i$) obviously do commute; that is,

$$\mathfrak{p}_j\mathfrak{q}_i - \mathfrak{q}_i\mathfrak{p}_j = 0. \tag{106'}$$

Also evidently, two $\mathfrak{q}$ or two $\mathfrak{p}$ commute:

$$\mathfrak{q}_j\mathfrak{q}_i - \mathfrak{q}_i\mathfrak{q}_j = 0, \tag{107}$$

$$\mathfrak{p}_j\mathfrak{p}_i - \mathfrak{p}_i\mathfrak{p}_j = 0. \tag{108}$$

Relations (106) and (106′), of fundamental importance in quantum mechanics, are known as the *commutation relations.*

In quantum mechanics an observable and its operator are often denoted by the same symbol (or the corresponding matrix), rather than being distinguished by different letters as has been done here. This practice leads to writing the commutation relations (106) in the form

$$p_iq_i - q_ip_i = \frac{h}{2\pi i}.$$

The paradoxcail aspect of these equations disappears if we recall that they do not refer to the physical quantities p_i and q_i but merely to their operators or corresponding matrices.

Let us now establish some additional important commutation relations. Let $Q(q_1, q_2, \dots q_f)$ be a function of the q alone, and let us consider it as an operator $\mathfrak{Q} = Q(q_1, q_2, \dots q_f)$. For any (scalar) function f we then have

$$\mathfrak{p}_i\mathfrak{Q}f = \frac{h}{2\pi i}\frac{\partial}{\partial q_i}(Qf) = \frac{h}{2\pi i}\frac{\partial Q}{\partial q_i}f + \frac{h}{2\pi i}Q\frac{\partial f}{\partial q_i},$$
$$\mathfrak{Q}\mathfrak{p}_i f = \frac{h}{2\pi i}Q\frac{\partial f}{\partial q_i},$$

and hence we may write the commutation relation

$$\mathfrak{p}_i\mathfrak{Q} - \mathfrak{Q}\mathfrak{p}_i = \frac{h}{2\pi i}\frac{\partial Q}{\partial q_i}. \tag{109}$$

For a (rational integral) function $P(p_1, p_2, \dots p_f)$ of the p alone, an analogous relation holds: calling $\mathfrak{P}$ the operator $P(\mathfrak{p}_1, \mathfrak{p}_2, \dots \mathfrak{p}_f)$, we obtain

$$\mathfrak{P}\mathfrak{q}_i - \mathfrak{q}_i\mathfrak{P} = \frac{h}{2\pi i}\frac{\partial\mathfrak{P}}{\partial\mathfrak{p}_i}, \tag{110}$$

where $\partial\mathfrak{P}/\partial\mathfrak{p}_i$ is to designate the operator which is obtained from the function $\partial P/\partial p_i$ upon substitution of the operators $\mathfrak{p}_i$ for the p_i. This formula may be derived by observing that $\mathfrak{P}$ will be made up of terms of the type $a\mathfrak{p}_1^{n_1}\mathfrak{p}_2^{n_2} \dots \mathfrak{p}_f^{n_f}$, and therefore that it will suffice to prove (110) for an expression of this form. Now all these factors except $\mathfrak{p}_i^{n_i}$ commute with $\mathfrak{q}_i$; and for $\mathfrak{p}_i^{n_i}$ the following commutation rule holds [which may be found by successively applying (106) n_i times]:

$$\mathfrak{p}_i^{n_i}\mathfrak{q}_i - \mathfrak{q}_i\mathfrak{p}_i^{n_i} = n_i\frac{h}{2\pi i}\mathfrak{p}_i^{n_i-1} = \frac{h}{2\pi i}\frac{\partial\mathfrak{p}_i^{n_i}}{\partial\mathfrak{p}_i}.$$

Thus (110) is proved.

From (109) and (110) we obtain, for a function G of the form $G = P(p) + Q(q)$, the following commutation relations (which contain all the previously cited cases as particular instances):

$$\mathfrak{p}_i\mathfrak{G} - \mathfrak{G}\mathfrak{p}_i = \frac{h}{2\pi i}\frac{\partial\mathfrak{G}}{\partial\mathfrak{q}_i}, \tag{111}$$
$$\mathfrak{G}\mathfrak{q}_i - \mathfrak{q}_i\mathfrak{G} = \frac{h}{2\pi i}\frac{\partial\mathfrak{G}}{\partial\mathfrak{p}_i}. \tag{112}$$

27. Average values. Quantum mechanics is designed to deal with so-called *elementary phenomena*, namely, those in which only one or a small number of elementary particles are involved. However, it is important to show that when quantum mechanics is applied to *macroscopic* systems—that is, systems composed of a very large number of elementary particles, such as ordinary bodies—its results coincide with the results of classical mechanics. Therefore classical mechanics is not in conflict with quantum mechanics but simply represents a limiting case of the latter.

We shall call "macroscopic observation" an observation which is equivalent to the measurement of the same observable G on a large number of equal systems (such as atoms) and to taking the average value of the results obtained. The greater part of ordinary physical measurements are just of this type. For example, the determination of an electric current is equivalent to taking the average of the velocities of the single electrons constituting the current (and to multiplying it by an appropriate factor); the determination of the coordinates of the center of mass of a body is equivalent to measuring the coordinates of the separate molecules and to taking their average; and so forth. From this introduction it may be seen why the study of the "average value" of an observable in an ensemble of systems (of identical constitution) is of interest.

If all the systems of the ensemble under consideration are in the same "state," the ensemble is said to represent a *pure case;* otherwise it is called a *mixture* and may be decomposed into a number of partial ensembles, each of which represents a pure case. In what is to follow, we shall always refer to a pure case unless a mixture is specified.

If in each system of the ensemble the measurement of the observable G can give the results G_r, with respective probabilities P_r, the average value $\bar{G}$ of all the results obtained when measuring G in all the systems of the ensemble will be

$$\bar{G} = \sum_r G_r P_r. \tag{113}$$

Now if the state of each system is represented by the vector ψ (which is the same for all systems) and if we call φ_r the eigenfunction of the operator $\mathfrak{G}$ corresponding to the eigenvalue[20] G_r, we have

$$\psi_r = \psi \cdot \varphi_r, \tag{114}$$

and the projection of ψ upon the vector φ_r is $P_r = \psi_r\psi_r^*$, by the fundamental principle of quantum mechanics, and hence

$$\bar{G} = \sum_r G_r\psi_r\psi_r^*.$$

We may transform this formula further by using the fact that $\psi_r^* = \varphi_r \cdot \psi$, according to (114). Furthermore, since $\mathfrak{G}\varphi_r = G_r\varphi_r$ and since ψ_r is a constant, we have

$$\bar{G} = \sum_r G_r\psi_r(\varphi_r \cdot \psi) = \sum_r \psi_r(\mathfrak{G}\varphi_r) \cdot \psi = \left(\mathfrak{G} \sum_r \psi_r\varphi_r\right) \cdot \psi,$$

or

$$\bar{G} = (\mathfrak{G}\psi) \cdot \psi, \tag{115}$$

or else in explicit form

$$\bar{G} = \int \psi^*\mathfrak{G}\psi \, dS, \tag{115'}$$

where, as usual, $dS = dq_1\, dq_2 \, . \, . \, . \, dq_f$, and the integral is understood to be extended over the whole space of the q. As may be seen, to a definite *state* ψ of the systems there corresponds a definite *average value* for any observable; hence *for macroscopic observations there is no uncertainty principle.*

If the ensemble is a mixture, it will be decomposed into partial ensembles, in each of which the state of the systems will be represented by a vector ψ^j. The average value $\bar{G}^j$ for each of these is to be calculated by means of (115). The required average value will be

$$\bar{G} = \sum_j c^j\bar{G}^j, \tag{116}$$

where $c^j = N^j/N$ is the ratio of the number N^j of systems in the state ψ^j to the total number N.

If G is a Cartesian coordinate q_i, relation (115') yields

$$\bar{q}_i = \int q_i\psi^*\psi \, dS; \tag{116'}$$

---

[20] In order to include cases of degeneracy also, it is necessary to make p separate terms correspond to each multiple eigenvalue of order p in (113).

and if G is a Cartesian component of momentum p_i,

$$\bar{p}_i = \frac{h}{2\pi i} \int \psi^* \frac{\partial \psi}{\partial q_i} \, dS. \tag{117}$$

These formulas are identical with those of Chapter 5, which define the center of a wave packet and its average propagation vector.

28. Time derivatives. Classical mechanics as a limiting case. Thus far we have considered an observable to be defined only for a certain instant $t = t_0$. However, it is obviously appropriate to introduce something which is analogous to the classical concept of "physical quantity as a function of time." Let us therefore consider that the same physical process which, when put into effect at the time t_0, defines the observable G_{t_0}, will, when put into effect at another instant t, define another observable G_t. In this way we are led to consider a continuous succession of observables which is entirely analogous to the continuous succession of the values of a physical quantity as a function of time in classical mechanics. We shall sometimes say, by analogy, that we are dealing with an "observable which is a function of t." A general observable G_t will be represented by the same operator $\mathfrak{G}(\mathfrak{q}, \mathfrak{p})$ which represents G_{t_0}, since the physical process defining it is the same. However, by supposing that some of the parameters which enter into the definition of the process of observation are functions of t, we shall also be able to generalize the concept of variation in time. Then the operator will contain t as well as the $\mathfrak{q}$ and $\mathfrak{p}$; that is, it will depend explicitly on the time (it will then be said that the observable G depends *explicitly* on t). In that case, the eigenvalues and the eigenfunctions of G will in general be functions of t. In the case that G does not contain t explicitly, the eigenvalues are constants, and the eigenfunctions are independent of t; that is, the principal axes are fixed in Hilbert space.

We shall now see how to introduce for the observables the analogue of the time derivative of a classical physical quantity. It is clear that one may *not* define, for the derivative $\dot{G}$ of the observable G, the expression

$$\lim_{dt \to 0} \frac{G_{t+dt} - G_t}{dt},$$

since this way of writing has no meaning; G_{t+dt} and G_t do not represent two numbers but two procedures of observation referring to different instants, each of which may yield an infinite number of numerical results. Instead, it is necessary to have recourse to a consideration of the average values and to define $\dot{G}$ as an observable such that its average value (defined as in the preceding section) is equal to the derivative of the average value of the observable G.[21]

Let us therefore differentiate (115), supposing, for greater generality, that the operator $\mathfrak{G}$ depends explicitly on t. We have

$$\dot{\bar{G}} = \frac{\partial}{\partial t}(\mathfrak{G}\psi)\cdot\psi + (\mathfrak{G}\psi)\cdot\frac{\partial\psi}{\partial t}$$

and also, indicating by $\partial\mathfrak{G}/\partial t$ the operator obtained from the expression for $\mathfrak{G}$ by the formal operation of differentiation with respect to t (hence $\partial\mathfrak{G}/\partial t = 0$ if $\mathfrak{G}$ does not contain t explicitly),

$$\dot{\bar{G}} = \left(\frac{\partial\mathfrak{G}}{\partial t}\psi + \mathfrak{G}\frac{\partial\psi}{\partial t}\right)\cdot\psi + \mathfrak{G}\psi\cdot\frac{\partial\psi}{\partial t}.$$

Substituting expression (87) for the derivative of ψ, we have, recalling (5′),

$$\dot{\bar{G}} = \left(\frac{\partial\mathfrak{G}}{\partial t}\psi\right)\cdot\psi - \frac{2\pi i}{h}(\mathfrak{G}\mathfrak{H}\psi)\cdot\psi + \frac{2\pi i}{h}(\mathfrak{G}\psi)\cdot(\mathfrak{H}\psi).$$

But since $\mathfrak{H}$ is Hermitian, we have (see §9)

$$(\mathfrak{G}\psi)\cdot(\mathfrak{H}\psi) = (\mathfrak{H}\mathfrak{G}\psi)\cdot\psi,$$

---

[21] This amounts to saying that from an ensemble of a very large number of samples of the system as defined in the preceding section (possibly a mixture), we select at random a certain (rather large) number of systems, observe G_t on them, and find the average value. Then we calculate the numerical quantity

$$\lim_{dt\to 0}\frac{\bar{G}_{t+dt} - \bar{G}_t}{dt}.$$

Finally we postulate that there exists an observable $\dot{G}$ such that its average value (taken from a third group of samples removed from the same ensemble) is equal to that quantity, no matter what the composition of the ensemble. Thus instead of measuring G_t and G_{t+dt} on the same system (a method which gives rise to the difficulty that the second measurement is carried out on the system in a different state from the one before), we perform the two measurements on different systems.

and hence

$$\dot{\bar{G}} = \left[\frac{\partial \mathfrak{G}}{\partial t} + \frac{2\pi i}{h}(\mathfrak{H}\mathfrak{G} - \mathfrak{G}\mathfrak{H})\right]\psi \cdot \psi.$$

Comparing this with (115), we note that $\dot{\bar{G}}$ is calculated formally as the average value of an observable whose operator is the expression contained in the square brackets. We shall therefore represent the derivative of the observable G symbolically by the formula

$$\dot{G} = \frac{\partial G}{\partial t} + \frac{2\pi i}{h}(\mathcal{H}G - G\mathcal{H}), \tag{118}$$

to which corresponds the analogous relation between the operators, where the operator corresponding to the observable $\dot{G}$ is denoted by $\dot{\mathfrak{G}}$,

$$\dot{\mathfrak{G}} = \frac{\partial \mathfrak{G}}{\partial t} + \frac{2\pi i}{h}(\mathfrak{H}\mathfrak{G} - \mathfrak{G}\mathfrak{H}). \tag{118'}$$

If G does not depend explicitly on t, the first term on the right hand side will be absent.

Here we may give a meaningful interpretation to the derivative $\dot{G}$ defined by (118), or, better, to the differential $\dot{G}\,dt$. Consider the observable $G + \dot{G}\,dt$, which we shall call g, and suppose that its measurement (at time t) yields the value g'. We shall show that after such an observation the system is left in a state such that if we measure the observable $\dot{G}$ at time $t + dt$, we find the value g'. Thus we may say that the measurement (at time t) of the observable $G + \dot{G}\,dt$ serves to determine the value which G *will have* at the time $t + dt$. It is to be noted, however, that such a measurement is incompatible with a measurement of G at time t. In order to prove this assertion, let us call ψ the function representing the state in which the system is left after a measurement of g has yielded the result g'. We shall have

$$\mathfrak{g}\psi(t) = g'\psi(t), \tag{119}$$

or, simply writing ψ for $\psi(t)$,

$$\mathfrak{G}\psi + \left[\frac{\partial \mathfrak{G}}{\partial t} + \frac{2\pi i}{h}(\mathfrak{H}\mathfrak{G} - \mathfrak{G}\mathfrak{H})\right]\psi\,dt = g'\psi. \tag{119'}$$

We are to verify that if we let this ψ evolve for a time dt, we shall obtain a $\psi(t + dt)$ which is an eigenfunction of the operator $\mathfrak{G}_{t+dt}$ corresponding to the eigenvalue g'—namely, that

$$\mathfrak{G}_{t+dt}\psi(t+dt) = g'\psi(t+dt),$$

or

$$\left(\mathfrak{G} + \frac{\partial \mathfrak{G}}{\partial t}\,dt\right)\left(\psi + \frac{\partial \psi}{\partial t}\,dt\right) = g'\left(\psi + \frac{\partial \psi}{\partial t}\,dt\right).$$

Expanding, and substituting for the derivative of ψ the value given by the Schrödinger equation (87), we have

$$\mathfrak{G}\psi + \left(\frac{\partial \mathfrak{G}}{\partial t} - \frac{2\pi i}{h}\,\mathfrak{G}\mathfrak{H}\right)\psi\,dt = g'\psi - \frac{2\pi i}{h}\,\mathfrak{H}g'\psi\,dt,$$

and, because of (119),

$$\mathfrak{G}\psi + \left(\frac{\partial \mathfrak{G}}{\partial t} - \frac{2\pi i}{h}\,\mathfrak{G}\mathfrak{H}\right)\psi\,dt = g'\psi - \frac{2\pi i}{h}\,\mathfrak{H}\mathfrak{g}\psi\,dt.$$

In the last term $\mathfrak{g}$ may be replaced by $\mathfrak{G}$ (up to terms in dt^2). The equation will then coincide with (119′), and thus is verified.

Let us apply this result to obtain again, in a generalized and more precise form, the result that a wave packet moves like a point particle in ordinary mechanics. In Part II this principle (limited to wave packets sufficiently small to be considered point-sized) constituted our point of departure for the establishment of the Schrödinger equation.

Identifying G with a coordinate q_i, we get from (118′)

$$\dot{\mathfrak{q}}_i = \frac{2\pi i}{h}\,(\mathfrak{H}\mathfrak{q}_i - \mathfrak{q}_i\mathfrak{H});$$

and identifying G with the momentum p_i, we obtain

$$\dot{\mathfrak{p}}_i = \frac{2\pi i}{h}\,(\mathfrak{H}\mathfrak{p}_i - \mathfrak{p}_i\mathfrak{H}).$$

Since the Hamiltonian $\mathcal{H}$ is of the form $P(p) + Q(q)$, we may apply (111) and (112), finding

$$\dot{\mathfrak{q}}_i = \frac{\partial \mathfrak{H}}{\partial \mathfrak{p}_i}, \qquad \dot{\mathfrak{p}}_i = -\frac{\partial \mathfrak{H}}{\partial \mathfrak{q}_i}.$$

According to what was said in connection with expression (118), these relations between operators stand for the following relations between the average values of the corresponding observables:

$$\dot{\bar{q}}_i = \overline{\frac{\partial \mathcal{H}}{\partial p_i}}, \qquad \dot{\bar{p}}_i = -\overline{\frac{\partial \mathcal{H}}{\partial q_i}}. \tag{120}$$

Hence, Hamilton's equations hold *on the average.* For instance, for a point particle in Cartesian coordinates, we have

$$\mathcal{H} = \frac{1}{2m}\sum_{i=1}^{3} p_i^2 + U(q),$$

and hence the preceding equations yield

$$\dot{\bar{q}}_i = \frac{1}{m}\,\bar{p}_i, \qquad \dot{\bar{p}}_i = -\overline{\frac{\partial U}{\partial q_i}} = \bar{F}_i,$$

where $\bar{F}_i$ denotes the components of the force. Eliminating p_i from these equations, we obtain

$$m\ddot{\bar{q}}_i = \bar{F}_i; \tag{121}$$

that is: the center of the wave packet moves like a material point obeying classical mechanics, subject to a force which is calculated by taking the average value of the force on the entire wave packet (Ehrenfest's theorem). It is apparent that this theorem applies to a wave packet of any size, whereas in the considerations which led to the Schrödinger equation in Part II, we referred to the limiting case of a practically point-sized packet. Furthermore, by applying (116), it may easily be seen that the theorem holds also for the average values of a mixture.

29. First integrals. In classical mechanics the *first integral* of a problem is an expression $G(q, p)$ of such a nature that it reduces to a constant if the q and the p vary in time so as to satisfy the equations of dynamics.

Analogously, in quantum mechanics we define as first integral an observable G such that its derivative $\dot{G}$ defined by (118) is identically zero, that is, such that

$$\frac{\partial \mathfrak{G}}{\partial t} + \frac{2\pi i}{h}(\mathfrak{H}\mathfrak{G} - \mathfrak{G}\mathfrak{H}) = 0. \tag{122}$$

It may readily be seen that the energy $\mathcal{H}$ is a first integral if (and only if) $\partial\mathcal{H}/\partial t = 0$, that is, if the Hamiltonian does not contain the time explicitly. In that case, the system is said to be *conservative.*

It may be shown that if G is a first integral, its eigenvalues will be constants (even if G contains the time t explicitly), and furthermore, although the principal axes of $\mathfrak{G}$ are not fixed, their rotation in Hilbert space bears a relation to the rotation of the state vector ψ (no matter what ψ may be) such that the projections of ψ upon these axes maintain a constant value. This means that the probabilities of the single eigenvalues do not vary in time.[22]

Generally we shall confine ourselves to a consideration of first integrals which do not contain t explicitly. Then (122) becomes

$$\mathfrak{H}\mathfrak{G} = \mathfrak{G}\mathfrak{H},$$

which expresses the fact that *the necessary and sufficient condition for an observable G (not containing the time t) to be a first integral is that its operator commute with the Hamiltonian, or that the observation of G be compatible with a simultaneous observation of the energy.*

30. Angular momenta and their operators. As an application of the preceding sections, let us consider the three observables M_x, M_y, M_z, the *angular momenta* of a particle with respect to the x-, y-, and z-axes, and let us first of all find the operators $\mathfrak{M}_x$, $\mathfrak{M}_y$, $\mathfrak{M}_z$ corresponding to them.

For instance, considering M_z, we observe that its expression in classical mechanics is

$$M_z = xp_y - yp_x. \tag{123}$$

Hence, according to the rule of §22, the operator corresponding to this expression is

$$\mathfrak{M}_z = x\mathfrak{p}_y - y\mathfrak{p}_x = \frac{h}{2\pi i}\left(x\frac{\partial}{\partial y} - y\frac{\partial}{\partial x}\right), \tag{124}$$

and similarly for M_x and M_y. First we note that any two of these observables are incompatible.[23] We have, for instance, that

$$\mathfrak{M}_x\mathfrak{M}_y = (y\mathfrak{p}_z - zp_y)(z\mathfrak{p}_x - x\mathfrak{p}_z) = y\mathfrak{p}_x\mathfrak{p}_z z - yx\mathfrak{p}_z^2 - z^2\mathfrak{p}_y\mathfrak{p}_x + zx\mathfrak{p}_y\mathfrak{p}_z,$$

$$\mathfrak{M}_y\mathfrak{M}_x = x\mathfrak{p}_y\mathfrak{p}_z z - xy\mathfrak{p}_z^2 - z^2\mathfrak{p}_x\mathfrak{p}_y + zy\mathfrak{p}_x\mathfrak{p}_z;$$

---

[22] For a proof, see, for example, No. 4 of the Bibliography, page 229.

[23] This incompatibility is the cause of the paradoxical nature of the angular momentum quantization in the Bohr-Sommerfeld theory (cf. §56 and §62). In fact, M_x, M_y, M_z cannot be considered as the components of an ordinary vector since, by reason of the conceptual impossibility of measuring two of them simultaneously, the direction of this vector turns out to be conceptually unobservable.

and hence, because of (106) and (124),

$$\mathfrak{M}_x\mathfrak{M}_y - \mathfrak{M}_y\mathfrak{M}_x = (y\mathfrak{p}_x - x\mathfrak{p}_y)(\mathfrak{p}_z z - z\mathfrak{p}_z) = -\frac{h}{2\pi i}\mathfrak{M}_z.$$

Therefore the following commutation relations hold for angular momenta:

$$\left.\begin{aligned} \mathfrak{M}_y\mathfrak{M}_z - \mathfrak{M}_z\mathfrak{M}_y &= -\frac{h}{2\pi i}\mathfrak{M}_x, \\ \mathfrak{M}_z\mathfrak{M}_x - \mathfrak{M}_x\mathfrak{M}_z &= -\frac{h}{2\pi i}\mathfrak{M}_y, \\ \mathfrak{M}_x\mathfrak{M}_y - \mathfrak{M}_y\mathfrak{M}_x &= -\frac{h}{2\pi i}\mathfrak{M}_z. \end{aligned}\right\} \tag{125}$$

When vector notation is introduced for the operators, the above relations may be expressed by the symbolic formula

$$\mathfrak{M} \times \mathfrak{M} = -\frac{h}{2\pi i}\mathfrak{M}. \tag{125'}$$

Let us now look for the eigenfunctions and eigenvalues of these operators. Let us take $\mathfrak{M}_z$, for example. We note that upon introducing polar coordinates r, θ, φ, with the z-axis as polar axis, we obtain

$$\frac{\partial}{\partial\varphi} = x\frac{\partial}{\partial y} - y\frac{\partial}{\partial x}$$

and hence (124) becomes

$$\mathfrak{M}_z = \frac{h}{2\pi i}\frac{\partial}{\partial\varphi}. \tag{126}$$

Therefore the equation for the eigenfunctions $\psi(r, \theta, \varphi)$ becomes

$$\frac{h}{2\pi i}\frac{\partial\psi}{\partial\varphi} = M'_z\psi,$$

where M'_z stands for a general eigenvalue. Hence

$$\psi = f(r, \theta)\, e^{(2\pi i/h)M_z'\varphi},$$

where f is an arbitrary function of r and θ (its presence is due to the fact that $\mathfrak{M}_z$ is an incomplete operator). This ψ is evidently periodic in φ with period h/M'_z. However, for ψ to have only one value at each point of space, it must be periodic in φ with period 2π. Hence we must have $mh/M'_z = 2\pi$ with m an integer, or

$$M'_z = m\frac{h}{2\pi}.$$

These are the desired eigenvalues; the same would be found for M_x and M_y.

Therefore the result of a measurement of the angular momentum with respect to an axis is always a (positive, zero, or negative) multiple of $h/2\pi$. This result extends and clarifies the "spatial quantization" of the Sommerfeld theory (see §56 of Part II).

We may also readily see that M_z is (as in classical mechanics) a first integral if the forces have zero moment with respect to the z-axis. As a matter of fact, in that case the potential U, expressed in polar coordinates, must be independent of φ; hence the Hamiltonian $\mathfrak{H}$ does not contain φ, and therefore commutes with the operator (126).

Let us now consider the observable M, the modulus of the angular momentum of a particle with respect to the origin. Classically we have $M^2 = M_x^2 + M_y^2 + M_z^2$. Hence for the operator corresponding to M^2 we shall take

$$\mathfrak{M}^2 = \mathfrak{M}_x^2 + \mathfrak{M}_y^2 + \mathfrak{M}_z^2, \tag{127}$$

where $\mathfrak{M}_x$, and so on, are given by (124) and analogous expressions. Inserting these and taking (106) into account, we find, by a simple calculation,

$$\mathfrak{M}^2 = r^2 \sum \mathfrak{p}_x^2 - \left(\sum x\mathfrak{p}_x\right)^2 - \frac{h}{2\pi i}\sum x\mathfrak{p}_x,$$

where Σ indicates a sum carried out by successively changing x into y and z. Now, recalling the meaning of the operators $\mathfrak{p}_x$, $\mathfrak{p}_y$, $\mathfrak{p}_z$, we see that

$$\sum \mathfrak{p}_x^2 = -\frac{h^2}{4\pi^2}\Delta,$$
$$\sum x\mathfrak{p}_x = \frac{h}{2\pi i} r\frac{\partial}{\partial r},$$

and hence

$$\mathfrak{M}^2 = -\frac{h^2}{4\pi^2}\left[r^2\Delta - r\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}\right) - r\frac{\partial}{\partial r}\right]. \tag{128}$$

We now recall that the Laplacian operator Δ in polar coordinates is given by

$$\Delta = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial}{\partial r}\right) + \frac{1}{r^2} \quad , \tag{129}$$

where we have denoted by Λ the operator, independent of r,

$$\Lambda = \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\varphi^2}. \tag{130}$$

Substituting into (128), we find

$$\mathfrak{M}^2 = -\frac{h^2}{4\pi^2}\Lambda. \tag{131}$$

The values of the observable M^2 are therefore given by $-(h^2/4\pi^2)\Lambda'$, where Λ' is an eigenvalue of the equation

$$\Lambda\psi = \Lambda'\psi. \tag{132}$$

This equation is none other than (223′) of §46, Part II—the differential equation for spherical harmonics (Λ' corresponding to $-C$). As we have seen, its eigenvalues are $\Lambda' = -l(l+1)$, with $l = 0$, 1, 2, ... Therefore the values which the observable M, or angular momentum, may take are given by

$$M = \sqrt{l(l+1)}\,\frac{h}{2\pi}. \tag{133}$$

This result has been stated in §46 of Part II.

It is to be noted that the operator $\mathfrak{M}^2$, and hence also $\mathfrak{M}$, commutes with each of the operators $\mathfrak{M}_x$, $\mathfrak{M}_y$, $\mathfrak{M}_z$. In fact, we have, because of (125),

$$\begin{aligned}
\mathfrak{M}^2\mathfrak{M}_x - \mathfrak{M}_x\mathfrak{M}^2 &= \mathfrak{M}_y^2\mathfrak{M}_x - \mathfrak{M}_x\mathfrak{M}_y^2 + \mathfrak{M}_z^2\mathfrak{M}_x - \mathfrak{M}_x\mathfrak{M}_z^2 \\
&= \mathfrak{M}_y\left(\mathfrak{M}_x\mathfrak{M}_y + \frac{h}{2\pi i}\mathfrak{M}_z\right) - \left(\mathfrak{M}_y\mathfrak{M}_x - \frac{h}{2\pi i}\mathfrak{M}_z\right)\mathfrak{M}_y \qquad (134) \\
&\quad + \mathfrak{M}_z\left(\mathfrak{M}_x\mathfrak{M}_z - \frac{h}{2\pi i}\mathfrak{M}_y\right) - \left(\mathfrak{M}_z\mathfrak{M}_x + \frac{h}{2\pi i}\mathfrak{M}_y\right)\mathfrak{M}_z = 0;
\end{aligned}$$

similarly we show that $\mathfrak{M}^2$ commutes with $\mathfrak{M}_y$ and $\mathfrak{M}_z$. Thus the measurement of the total angular momentum is compatible with the measurement of its projection upon any direction, whereas the measurements of two components of the angular momentum are incompatible.

We recognize, then, that M^2 is a first integral (as in classical mechanics), if the force is of the central type. In fact, the operator $\mathfrak{H}$ for a particle is (see §19)

$$\mathfrak{H} = -\frac{h^2}{8\pi^2 m}\Delta + U = -\frac{h^2}{8\pi^2 m}\left[\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right) + \frac{1}{r^2}\Lambda\right] + U; \qquad (135)$$

and since Λ is an operator which does not involve r, it always commutes with the first two terms of this expression. If, then, the force is central, U is a function of r alone, and hence Λ commutes also with the last term. In that case, therefore, Λ commutes with $\mathfrak{H}$, and similarly, by (131), for $\mathfrak{M}^2$, so that M^2 is a first integral.

31. Magnetic forces. In all the considerations thus far we have dealt only with particles acted upon by forces derivable from a potential U. The physically rather important case of electrically charged particles moving in a magnetic field does not fall within the preceding considerations and therefore requires an additional extension of the latter. This extension, like the previous ones, will be carried out by using the considerations of §19 as a guide.

If a particle of charge e moves with velocity $\mathbf{v}$ in an electric field $\mathbf{E}$ and in a magnetic field $\mathbf{H}$, it will experience a force

$$\mathbf{F} = e\left(\mathbf{E} + \frac{1}{c}\mathbf{v} \times \mathbf{H}\right), \qquad (136)$$

whose electric part is derivable from a potential U (putting $e\mathbf{E} = -\text{grad } U$), whereas the magnetic part does not have a potential. We note that the ordinary dynamical equations of a particle may in this case also be written in the Hamiltonian form, if the *vector potential* $\mathbf{A}$ is introduced[24] and the function

$$\mathfrak{H} = \sum_{r=1}^{3} \frac{1}{2m}\left(p_r - \frac{e}{c}A_r\right)^2 + U(q_1, q_2, q_3) \qquad (138)$$

is taken for the Hamiltonian.

---

[24] We recall that the electric field $\mathbf{E}$ and the magnetic field $\mathbf{H}$ are derived from the scalar potential V and vector potential $\mathbf{A}$ by the known formulas

$$\mathbf{E} = -\text{grad } V - \frac{1}{c}\frac{\partial \mathbf{A}}{\partial t}, \qquad \mathbf{H} = \text{curl } \mathbf{A}. \qquad (137)$$

It is further to be remembered that the potential U used here is equal to eV.

In fact, Hamilton's equations as obtained from this expression are

$$\dot{q}_s = \frac{\partial \mathcal{H}}{\partial p_s} = \frac{1}{m}\left(p_s - \frac{e}{c}A_s\right)$$

$$\dot{p}_s = -\frac{\partial \mathcal{H}}{\partial q_s} = \frac{e}{c}\sum_{r=1}^{3}\frac{1}{m}\left(p_r - \frac{e}{c}A_r\right)\frac{\partial A_r}{\partial q_s} - \frac{\partial U}{\partial q_s}.$$

The first equation yields[25]

$$p_s = m\dot{q}_s + \frac{e}{c}A_s \tag{139}$$

and the second one, through use of equations (137) and (136), yields the equation of dynamics

$$m\ddot{q}_s = eF_s.$$

Since we want to preserve the analogy mentioned in §19, when a magnetic field exists we shall obtain the equation of wave mechanics by transforming the Hamiltonian (138) into an operator by means of the usual substitution (*S*) (of page 322) that is, by putting

$$\mathfrak{H} = \sum_{r=1}^{3}\frac{1}{2m}\left(\frac{h}{2\pi i}\frac{\partial}{\partial q_r} - \frac{e}{c}A_r\right)^2 + U(q_j) \tag{140}$$

and writing the equation for ψ_n or for ψ in the usual form (81) or (82). We may also say that the operator $\mathfrak{H}$ corresponding to the presence of a magnetic field is obtained from the Hamiltonian with no field by substituting for p_r the operator

$$\mathfrak{P}_r = \mathfrak{p}_r - \frac{e}{c}A_r = \frac{h}{2\pi i}\frac{\partial}{\partial q_r} - \frac{e}{c}A_r. \tag{141}$$

The Schrödinger equation for stationary states is therefore, for a particle in a magnetic field,

$$\frac{1}{2m}\sum_{r=1}^{3}\left(\frac{h}{2\pi i}\frac{\partial}{\partial q_r} - \frac{e}{c}A_r\right)^2\psi_n = (E_n - U)\psi_n, \tag{142}$$

---

[25] Note the fact, expressed by this formula, that in the presence of a magnetic field the *momentum components* p_1, p_2, p_3 are no longer the velocity components multiplied by m. A *fixed* particle in a magnetic field has momentum components different from zero.

and the time-dependent equation for any state is

$$\frac{1}{2m}\sum_{r=1}^{3}\left(\frac{h}{2\pi i}\frac{\partial}{\partial q_r}-\frac{e}{c}A_r\right)^2\psi+U\psi=-\frac{h}{2\pi i}\frac{\partial\psi}{\partial t}. \qquad (142')$$

The justification for this extension of the Schrödinger equation is given by the following *theorem*, which we shall merely state here:[26] A sufficiently small wave packet constructed by means of the ψ_n satisfying (142) will move in the manner in which a particle of charge e would move under the action of forces derivable from the potential U, and of the magnetic field derivable from the vector potential **A**.

The above considerations may immediately be extended to a system of N distinct particles. In that case the Hamiltonian operator is, in the same notation as that of (84),

$$\mathcal{H}=\sum_{j=1}^{3N}\frac{1}{2m_j}\left(\frac{h}{2\pi i}\frac{\partial}{\partial q_j}-\frac{e_j}{c}A_j\right)^2+U(q_j),$$

where A_j denotes the component of the vector potential corresponding to the coordinate q_j.

32. Determinism and quantum mechanics.[27] In rational mechanics, as we know, the solution of the following problem is uniquely determined: given, at a certain instant $t = 0$, the positions and velocities of all the points of a system (subject to forces depending in a known way upon the positions and velocities), calculate the value of any coordinate or component of velocity of the system, or any function of these quantities, at any other instant t_1. This is an analytical property of the fundamental equations of mechanics, which, in the field of mechanics, gives precise expression to the general philosophical concept known as *determinism.*

Let us now state this problem from the viewpoint of quantum mechanics. As we have stated several times, no physical significance may be attributed to the expression "totality of the positions

---

[26] For the proof, see for instance No. 14 of the Bibliography, page 109.

[27] For a more thorough discussion of this argument, see for example No. 21 of the Bibliography; also A. Eddington, *Sur le problème du determinisme* (*Actualitiès Scient. et Ind.*), No. 112; Paris: Hermann, 1934.

and velocities of the points of a system at a given instant"; and hence the statement cited above, valid for classical mechanics, loses all meaning in quantum mechanics. In its stead, the following property holds.

Let us suppose that at time $t = 0$ we have carried out a maximum observation. Its results represent a complete description of the system at the instant considered (or, they define its "state" completely). As we have seen, the laws of quantum mechanics allow us to calculate,[28] for any instant t_1 (possibly the same instant) the possible results of the measurement of any observable G, and the respective probabilities. In general, these possible results constitute a (discrete or continuous) infinity, which is to be expressed by saying that there is a certain "indeterminacy" in the value of the observed quantity (an indeterminacy which, we repeat, does not arise from an imperfection in the measurements or from a lack of completeness of initial information, but rather from the very nature of elementary phenomena, such as photon scattering and particle collisions, to which the laws of quantum mechanics are to be applied).

---

[28] It may be useful to retrace the scheme of this calculation, which results from the preceding sections. Let $g_1(q, p)$, $g_2(q, p)$ . . . be the observables measured at time zero (constituting a maximum observation), and g_1, g_2, . . . the values found. The ψ which characterizes the state of the system is determined, for $t = 0$, by the equations

$$\mathfrak{g}_1\psi_0 = g_1\psi_0, \qquad \mathfrak{g}_2\psi_0 = g_2'\psi_0, \; . \; . \; . \; . \tag{143}$$

ψ then evolves in time as governed by the Schrödinger differential equation

$$\mathfrak{H}\psi = -\frac{h}{2\pi i}\frac{\partial\psi}{\partial t}, \tag{144}$$

which, together with the initial value ψ_0 given by (143), defines ψ at any time t, and in particular $\psi(t_1)$. Then we write the equation

$$\mathfrak{G}\varphi = G'\varphi; \tag{145}$$

its eigenvalues G' give the possible results of a measurement of G, and the respective probabilities are given by $|\psi(t_1) \cdot \varphi_n|^2$. Geometrically speaking, we would say that the observations at time zero define the initial position of the state vector ψ, whereas (144) governs the manner in which the latter evolves in time and hence permits the calculation of the state vector at the time t_1. By projecting it upon the principal axes of the operator $\mathfrak{G}$, given by (145), we obtain the required probability amplitudes.

However, for certain particular observables, a single value may exist with probability of unity,[29] all others having probability zero; that is, the observable *does have* a definite value at time t_1. In this case, quantum mechanics evidently allows us to calculate this definite value without any uncertainty, starting from the initial conditions (results of a maximum observation).

In order to make this reasoning clearer, let us consider the example of the linear harmonic oscillator (see §39 of Part II), and let us suppose that we have measured its energy at the time $t = 0$, finding a result E_n. This is a maximum observation, which completely determines the state of the system (in fact, its ψ will be the eigenfunction ψ_n of the Schrödinger equation). If then, at any time t_1 (possibly the same instant), we carry out an observation of position, that is, of the observable x, we may find any value between $-\infty$ and $+\infty$. All we can say a priori concerning the outcome of this measurement is that its probability is distributed according to the density $|\psi_n|^2$ (that is, according to the curves of Fig. 29). However, if instead of measuring the observable x, we measure at time t_1 the observable E (energy), we are certain of finding the value E_n again. There is thus no uncertainty whatever for this observable. Naturally, the same conclusion would be valid for any other observable compatible with E.

It is to be noted that the maximum observation to be carried out at time zero may be chosen with a large degree of arbitrariness, and these different ways of defining the state of the system are not equivalent, as was mentioned in §18. In fact, the observables having definite values are different in the various cases; this statement is true not only for observables relative to the time zero, but also for those relative to any instant t_1. In the case of the oscillator, for example, instead of E we could measure the coordinate x, the momentum p, or any function of these quantities $g(x, p)$. In that case, however, it would no longer be the energy which has an a priori determinable value at time t_1, but (possibly) another observable G. It may even be shown[30] that, given an observable $G(x, p)$ relative to a given instant t_1, it is always possible to find a

---

[29] To be exact, this condition occurs for all those observables whose operator has a principal axis in the direction of the state vector ψ at the instant t_1.

[30] Cf. E. Fermi, *Rend. Acc. Linc.* **XI**, series 6, 1st sem. 1930, page 980, or *Nuovo Cimento* **VII** (1930), page 361.

maximum observation to be carried out at time zero such that its result will make it possible—without indeterminacy—to calculate the value of G at the time t_1. However, if we wanted to calculate, without indeterminacy, the value of another observable G' (which might also be the same G at a different instant), we should have to perform a different maximum observation at time zero, incompatible with the former if G and G' are incompatible with each other. In conclusion, we can say that the state of the system at time t_1 may be determined, not only by a (complete) set of observations at the time t_1, but also by a (complete) set of observations at time zero. In other words, the *state* of the system at time t_1 is uniquely determined by the state at time zero. In quantum mechanics, however, this proposition has a meaning different from that of classical mechanics, since we have seen that the concept of "state" in quantum mechanics does not imply the totality of positions and velocities of the point particles, as it does in classical mechanics.

It is to be noted that the above considerations apply only to "elementary" phenomena, that is, to phenomena in which only a limited number of atoms, electrons, photons, and so on, intervene; whereas for "macroscopic" phenomena, into which there enter a very large number of elementary particles (that is, for the major part of commonly observed phenomena), the ordinary relations of determinism naturally hold. This is not in contradiction with quantum mechanics; in fact, it is one of its necessary consequences. Indeed, a "macroscopic" observation (see §27) is equivalent to a measurement of an average value on a very large number of elementary systems. But in quantum mechanics, the average value of any observable is uniquely determined when the *state* of the system is known, as was pointed out in §27. Hence determinism in the classical sense is valid for the average values.

CHAPTER 12

The Matrix Method

33. General remarks. In Chapter 10 we have seen that, having fixed any complete system of orthogonal functions y_n (in geometrical terminology, an orthogonal system of axes in Hilbert space), we may represent any vector of this space by its components f_n with respect to these axes. Similarly, every linear operator $\mathfrak{A}$ may be represented by a certain matrix $\{\mathfrak{A}\}$. Let us now apply this representation to the "state vector" ψ (with which we have dealt at length in the last chapter) and to the linear operators which operate upon them and which, as we have seen, correspond to the various observables. Hence we are now to consider the state of the system as defined by the aggregate of the (infinitely many) components of the vector ψ lying along the preassigned system of axes y_n, and to make a matrix, rather than an operator, correspond to every observable. In this way the algebraic relations between observables will now be translated into an equal number of relations between matrices. Therefore this method of dealing with problems of quantum mechanics is called the *matrix method;* it is of course entirely equivalent, from the theoretical standpoint, to the operator method and to the method of wave mechanics which we have used thus far. Practical reasons alone determine the preference of one or the other method in the various cases.

In order to understand better the respective positions of the two methods, we may compare the operator method with that method of rational mechanics which is purely vectorial without the use of systems of reference; the wave-mechanical method and the matrix method might be compared with the procedure of classical mechanics using a particular coordinate system. More precisely, in the case of wave mechanics, Hilbert space refers to that particular system of axes which we have called "continuous" in §2 (each specified by a group of values of the "coordinates" of the system), whereas in the matrix method we refer to a general system of discrete axes.

The matrix method, as pointed out in Part I, was invented by Heisenberg and was the first form taken by quantum mechanics. However, the point of view from which it was presented in the early days is considerably different from the one just mentioned, to which we shall adhere in the remainder of this chapter.

Let us now briefly outline the fundamental idea of this method.

First let us fix a system of reference axes in Hilbert space, selecting[1] a certain observable K and taking for our reference axes the principal axes of its operator $\mathfrak{K}$, that is, the directions specified by the eigenfunctions y_n of the equation

$$\mathfrak{K}y_n = K_n y_n;$$

we shall then say that the matrices we are using are referred to the "*K-representation.*" As we shall see, the $\mathcal{H}$-representation (where $\mathcal{H}$ is the energy) is of particular interest. In this scheme the directions of the axes of reference are evidently given by the eigenfunctions ψ_n of the Schrödinger equation.

Once the "representation" has been fixed, there corresponds, to every observable A, a Hermitian matrix

$$\{\mathfrak{A}\} = \left\| \begin{matrix} A_{11} & A_{12} & A_{13} \ldots \\ A_{21} & A_{22} & A_{23} \ldots \\ A_{31} & A_{32} & A_{33} \ldots \\ \ldots & \ldots & \ldots \ldots \end{matrix} \right\| \tag{146}$$

If we want to relate this representation to the operator method, we must remember (see §5) that the elements of this matrix are obtained from the operator $\mathfrak{A}$ (corresponding to the observable A according to the rule of §22) by the formula

$$A_{mn} = (\mathfrak{A}y_n) \cdot y_m = \int y_m^* \mathfrak{A} y_n \, dS. \tag{147}$$

These elements are the coefficients of the linear transformation which passes from the components of any vector f to the components of $\mathfrak{A}f$ or, briefly, of the transformation which expresses the effect of the operator. In the matrix method, however, we consider the A_{mn} as elements characteristic of the observable, without relating them to the expression (147).

---

[1] If the system has several degrees of freedom, it is to be understood that K represents a maximum observable, or a complete set of observables (see §18), and the y_n are the eigenfunctions common to all their operators.

We recall from §12 that, in particular, the matrix which in the K-representation stands for the observable K, that is, the same matrix which serves to *define* the representation, is a diagonal matrix

$$\{\mathfrak{K}\} = \begin{Vmatrix} K_1 & 0 & 0 & \dots \\ 0 & K_2 & 0 & \dots \\ 0 & 0 & K_3 & \dots \\ \dots & \dots & \dots & \dots \end{Vmatrix}$$

whose elements are simply the eigenvalues of the operator $\mathfrak{K}$ (see §10) and hence represent the possible results of a measurement of K. The elements A_{mn} of the matrix $\{\mathfrak{A}\}$, however, generally do not have any simple physical significance, except those of the principal diagonal. In fact, A_{nn} represents [see formula (115), §27] the average value of the observable A in the state specified by the vector y_n, that is, in the state defined by the value K_n of K. The elements off the main diagonal $A_{mn}(m \neq n)$, although not susceptible of simple interpretation, are nevertheless related in an important way to the statistics of the experimental results. In fact, the average value of any power A^k of A (in the state defined by K_n) is evidently given by the nth diagonal term of the matrix $\{\mathfrak{A}^k\}$, and in its calculation the nondiagonal elements of $\{\mathfrak{A}\}$ also enter.

Of particular interest are the elements of the three matrices $\{\mathfrak{X}\}$, $\{\mathfrak{Y}\}$, $\{\mathfrak{Z}\}$, representing the components of the electric moment of the system in the $\mathcal{H}$-representation, or the expressions

$$X_{mn} = \int \psi_m^* X \psi_n \, dS, \text{ and so on;}$$

in fact, as was mentioned in §32 of Part II, the radiation emitted (or absorbed) in the transition from the state m to the state n corresponds qualitatively to the radiation which would be emitted by an oscillator whose electric moment has the components X_{mn}, Y_{mn}, Z_{mn}, along the three axes.

34. Commutation relations. The algebraic relations between observables may be translated into relations of the same form between the matrices which represent them, it being understood, of course, that the operations of sum and product of matrices are defined by the rules of §6. In particular, between the matrices $\{\mathfrak{q}_k\}$ and $\{\mathfrak{p}_k\}$ representing the coordinates and momenta, the following commutation relations hold in any representation:

$$\{\mathfrak{p}_k\}\{\mathfrak{q}_k\} - \{\mathfrak{q}_k\}\{\mathfrak{p}_k\} = \frac{h}{2\pi i}\{1\}, \tag{148}$$

$$\{\mathfrak{p}_k\}\{\mathfrak{q}_l\} - \{\mathfrak{q}_l\}\{\mathfrak{p}_k\} = 0 \qquad (\text{for } k \neq l); \tag{148'}$$

and, for any k and l,

$$\{\mathfrak{q}_k\}\{\mathfrak{q}_l\} - \{\mathfrak{q}_l\}\{\mathfrak{q}_k\} = 0 \tag{149}$$

$$\{\mathfrak{p}_k\}\{\mathfrak{p}_l\} - \{\mathfrak{p}_l\}\{\mathfrak{p}_k\} = 0. \tag{150}$$

If then G is a function of the q and the p of the form

$$G = P(p) + Q(q)$$

(where P stands for a rational integral function, and Q for any function), a matrix $\{\mathfrak{G}\}$ will correspond to G for which the following commutation relations, which follow immediately from (111) and (112) of §26, will hold in any representation:

$$\{\mathfrak{p}_i\}\{\mathfrak{G}\} - \{\mathfrak{G}\}\{\mathfrak{p}_i\} = \frac{h}{2\pi i}\left\{\frac{\partial\mathfrak{G}}{\partial\mathfrak{q}_i}\right\}, \tag{151}$$

$$\{\mathfrak{G}\}\{\mathfrak{q}_i\} - \{\mathfrak{q}_i\}\{\mathfrak{G}\} = \frac{h}{2\pi i}\left\{\frac{\partial\mathfrak{G}}{\partial\mathfrak{p}_i}\right\}. \tag{152}$$

In (151) and (152), the preceding relations are contained as special cases.

35. Search for eigenvalues by the matrix method. After these preliminary remarks, we shall see how the problem of finding the eigenvalues of an observable G (which, in particular, may be the energy) presents itself in the matrix method.

We must start, as in §22, from the analytic expression for the observable G as a function of the q and the p. This expression takes the place of a definition of G and, as has been mentioned, is usually constructed by analogy to classical mechanics. Sometimes it is necessary to remember that we must "symmetrize" the products p_kq_k in order that the matrix corresponding to G be Hermitian. Once this expression $G(q, p)$ has been constructed, it may be interpreted as a relation among the matrices corresponding to the observables G, q, p:

$$\{\mathfrak{G}\} = G(\{\mathfrak{q}\}, \{\mathfrak{p}\}) \tag{153}$$

which is true in any representation. In particular, in the G-representation the same relation will hold, with the additional condition that the matrix $\{\mathfrak{G}\}$ be diagonal. Hence if we represent $\{\mathfrak{G}\}$ in that representation, the problem is to determine the elements of the matrices $\{\mathfrak{q}\}$ and $\{\mathfrak{p}\}$ in such a way that the latter satisfy the commutation relations (148) and (150), and furthermore that the matrix $\{\mathfrak{G}\}$ calculated by means of (153) be diagonal. After solution of this problem (for which no general methods can be given), the elements of the diagonal matrix $\{\mathfrak{G}\}$ will yield the required eigenvalues. We shall show an example of this procedure in the following section.

36. Application to the problem of the harmonic oscillator. Let us take the case of a harmonic oscillator of mass m and restoring force $-Kx$, treated by wave-mechanical methods in §39 of Part II, and let us try to find, by the matrix method, the values its energy may assume.

The expression for the energy as a function of x and $p = m\dot{x}$ (Hamiltonian) is, in analogy to classical mechanics,

$$\mathcal{H} = \frac{1}{2m} p^2 + \frac{1}{2} Kx^2.$$

We are to determine the elements of the matrices $\{\mathfrak{x}\}$ and $\{\mathfrak{p}\}$ (referred to the $\mathcal{H}$-representation) such that the commutation relation

$$\{\mathfrak{p}\}\{\mathfrak{x}\} - \{\mathfrak{x}\}\{\mathfrak{p}\} = \frac{h}{2\pi i}\{1\} \tag{154}$$

hold, and such that the matrix

$$\{\mathfrak{H}\} = \frac{1}{2m}\{\mathfrak{p}\}^2 + \frac{K}{2}\{\mathfrak{x}\}^2 \tag{155}$$

be diagonal. Translating these matrix equations into equations for the corresponding elements, and indicating by E_n the diagonal elements (H_{nn}) of the matrix $\{\mathfrak{H}\}$, that is, the desired eigenvalues, we have[2]

$$\sum_{r=0}^{\infty} (p_{jr}x_{rk} - x_{jr}p_{rk}) = \frac{h}{2\pi i}\delta_{jk}, \tag{156}$$

$$E_k\delta_{jk} = \frac{1}{2m}\sum_r p_{jr}p_{rk} + \frac{K}{2}\sum_r x_{jr}x_{rk}. \tag{157}$$

---

[2] In this problem, we shall number the rows and columns of the matrices from 0 rather than 1, in order to conform to the convention adopted in the wave-mechanical treatment of the same problem, in which we have numbered the eigenvalues E_0, E_1, and so on.

Then applying to the matrix $\{\mathfrak{H}\}$ the commutation rules (151) and (152), we find

$$\{\mathfrak{p}\}\{\mathfrak{H}\} - \{\mathfrak{H}\}\{\mathfrak{p}\} = \frac{h}{2\pi i} K\{\mathfrak{x}\},$$

$$\{\mathfrak{x}\}\{\mathfrak{H}\} - \{\mathfrak{H}\}\{\mathfrak{x}\} = -\frac{h}{2\pi i}\frac{1}{m}\{\mathfrak{p}\}.$$

These equations may be translated into the following relations between the matrix elements:

$$\sum_r (p_{jr}H_{rk} - H_{jr}p_{rk}) = \frac{h}{2\pi i} Kx_{jk},$$

$$\sum_r (x_{jr}H_{rk} - H_{jr}x_{rk}) = -\frac{h}{2\pi i}\frac{1}{m} p_{jk},$$

or

$$p_{jk}(E_k - E_j) = \frac{h}{2\pi i} Kx_{jk}, \tag{158}$$

$$x_{jk}(E_k - E_j) = -\frac{h}{2\pi i}\frac{1}{m} p_{jk}, \tag{158'}$$

from which we get, upon eliminating p_{jk},

$$x_{jk}\left[1 - \frac{4\pi^2}{h^2}\frac{m}{K}(E_k - E_j)^2\right] = 0,$$

or else, upon introducing the classical frequency of the oscillator $\nu_0 = \frac{1}{2\pi}\sqrt{\frac{K}{m}}$,

$$x_{jk}\left[1 - \left(\frac{E_k - E_j}{h\nu_0}\right)^2\right] = 0. \tag{159}$$

From this it may be seen that the elements x_{jk} are all zero, except for those whose indices j and k are such that

$$E_k - E_j = \pm h\nu_0. \tag{160}$$

An analogous observation may be made for the elements p_{jk}. Therefore in the matrices $\{\mathfrak{x}\}$ and $\{\mathfrak{p}\}$ there are at most two elements different from zero in each row and each column.

Now the commutation relation (156) yields in particular, for a diagonal element ($j = k$),

$$\sum_r (p_{kr}x_{rk} - x_{kr}p_{rk}) = \frac{h}{2\pi i}.$$

Solving for the p in (158) and noting that $x_{kr} = x_{rk}^*$, we have

$$2K \sum_r \frac{|x_{kr}|^2}{E_r - E_k} = 1. \quad (k = 0, 1, 2, \ldots) \qquad (161)$$

From this expression we see that in each row of the matrix $\{\mathfrak{x}\}$ there is at least one nonzero element; otherwise, if the kth row consisted of all zero elements, the summation of (161) would be zero and the relation could not hold for this value of k. Hence if a certain number E_k is an eigenvalue, at least one of the quantities $E_k \pm h\nu_0$ is also an eigenvalue. This condition is satisfied if the eigenvalues form an arithmetic progression of constant difference $h\nu_0$, that is, when they are given by the formula

$$E_k = \epsilon + kh\nu_0, \qquad (162)$$

where ϵ is a constant which we shall determine presently. When the eigenvalues are numbered in this fashion, condition (160) is satisfied only for $j = k \pm 1$. Hence in the matrices $\{\mathfrak{x}\}$ and $\{\mathfrak{p}\}$, only the elements of the type $x_{k,k\pm1}$, $p_{k,k\pm1}$ are different from zero (these elements form two oblique lines, parallel and adjacent to the principal diagonal on either side). Expression (161) then yields, for $k = 0$,

$$2K \frac{|x_{01}|^2}{h\nu_0} = 1, \qquad (163)$$

and for $k = 1, 2, \ldots,$

$$2K \left[\frac{|x_{k,k+1}|^2}{h\nu_0} - \frac{|x_{k,k-1}|^2}{h\nu_0} \right] = 1, \qquad (164)$$

which may also be written

$$|x_{k,k+1}|^2 - |x_{k,k-1}|^2 = \frac{h\nu_0}{2K},$$

showing that the quantities $|x_{k,k+1}|^2$ form an arithmetic progression of constant difference $h\nu_0/2K$. The first term of that progression is given by (163), and is

$$|x_{01}|^2 = \frac{h\nu_0}{2K}. \tag{163'}$$

The general term is therefore

$$|x_{k,k+1}|^2 = (k+1)\frac{h\nu_0}{2K} = (k+1)\frac{h}{8\pi^2 m\nu_0}.$$

Consequently, for the elements $x_{k,k+1}$ of the matrix $\{\mathfrak{x}\}$ we may take[3]

$$x_{k,k+1} = \sqrt{k+1}\sqrt{\frac{h}{8\pi^2 m\nu_0}}. \tag{165}$$

We can deduce the elements of the form $x_{k,k-1}$ from these, observing that since the matrix must be Hermitian, $x_{k,k-1} = x^*_{k-1,k}$, and that this last quantity is obtained from (165) by simply changing k into $(k-1)$. Hence

$$x_{k,k-1} = \sqrt{k}\sqrt{\frac{h}{8\pi^2 m\nu_0}}. \tag{165'}$$

From these formulas by means of (158) or (158′) we obtain the expressions for the nonzero elements of the matrix $\{\mathfrak{p}\}$ as follows:

$$p_{k,k+1} = -i\sqrt{k+1}\sqrt{\frac{mh\nu_0}{2}}, \tag{166}$$

$$p_{k,k-1} = i\sqrt{k}\sqrt{\frac{mh\nu_0}{2}}. \tag{166'}$$

Thus the matrices $\{\mathfrak{x}\}$ and $\{\mathfrak{p}\}$ are completely determined. In order to find the energy levels, we have only to determine the constant ϵ of (162). This may be obtained immediately by writing (157) for the particular case $j = k = 0$. We have

$$E_0 = \frac{1}{2m}|p_{01}|^2 + \frac{K}{2}|x_{01}|^2,$$

---

[3] Of course we could add to these expressions a factor of the form $e^{i\delta_k}$, with δ_k arbitrary, but the eigenvalues would be the same, as may readily be recognized. This operation would correspond to a multiplication of the versors specifying the axes in Hilbert space, by factors of modulus 1, which do not alter anything of importance.

and taking $|x_{01}|^2$ and $|p_{01}|^2$ from (163′) and (166), we find

$$E_0 = \frac{h\nu_0}{2}.$$

Hence in (162) we must take $\epsilon = h\nu_0/2$, and the expression for the energy eigenvalues becomes

$$E_k = (k_k + \tfrac{1}{2})h\nu_0 \tag{167}$$

which coincides with the expression found by wave mechanics in §39 of Part II.

It may easily be verified that the expressions found for the elements of the matrices $\{\mathfrak{x}\}$ and $\{\mathfrak{p}\}$ also satisfy the relations (156) and (157) (which we have specialized by using only $j = k$) for $j \neq k$.

The elements of the matrices $\{\mathfrak{x}\}$ and $\{\mathfrak{p}\}$ which we have calculated (and which enter into the problems of radiation theory) could also be calculated by means of their wave-mechanical expression [see (147)]:

$$x_{jk} = \int \psi_j^* x \psi_k \, dx \tag{168}$$

$$p_{jk} = \frac{h}{2\pi i} \int \psi_j^* \frac{\partial \psi_k}{\partial x} \, dx, \tag{168′}$$

with the expressions found in §39, Part II, inserted for the eigenfunctions ψ. However, this procedure would lead to calculations considerably longer than those developed in this section. In the case of the oscillator, therefore, the matrix method presents certain advantages over the Schrödinger method.

CHAPTER 13

Perturbation Theory

37. General remarks. It is a well-known fact that in celestial mechanics the problem of the motion of the planets under the action of mutual attractions and of the attraction by the sun would be practically insoluble if all these forces were of the same order of magnitude. Fortunately, the attraction of the sun by far exceeds the attraction between the planets, so that the motion of a planet can be calculated by a method of successive approximations. This so-called "perturbation method" is capable of an accuracy which is more than sufficient for practical purposes. Essentially, it consists in first considering each planet to be subject to the solar attraction alone, under which assumption the problem is known to be rigorously soluble; then we calculate, by successive approximations, to what extent the motion is modified by the effect of the originally neglected mutual attractions (perturbing forces). An analogous procedure is followed in atomic physics in order to solve those problems in which the particles may be considered to be subject to certain predominant forces which, if they alone were present, would allow the complete solution of the problem ("unperturbed problem"), and subject also to other weaker "perturbing" forces. Thus, for instance, in the study of an atom located in a magnetic field (Zeeman effect), we can consider the action of the field upon the electrons of the atom as a perturbing force; supposing that we know how to solve the problem (that is, how to determine the eigenfunctions and eigenvalues) for the atom in the absence of the field, we can, by the method about to be described, determine the perturbation produced by the magnetic field on the eigenfunctions and eigenvalues.

Perturbation theory is of very great importance in atomic physics, since many problems of great experimental interest would lead, either by the method of wave mechanics or by the matrix method, to mathematical difficulties which would be either very acute or altogether insuperable, were it not for the perturbation method. This method may be adapted to the wave-mechanical as well as to the matrix approach, as we shall show in the following sections.

38. Perturbation of stationary states (nondegenerate case). First let us consider the unperturbed system, and let us call $\mathcal{H}^0(q, p)$ the Hamiltonian which describes it (we shall generally distinguish the quantities referring to the unperturbed problem by a superscript 0). Then the Schrödinger equation for the ith stationary state of the unperturbed system will be

$$\mathfrak{H}^0\psi_i^0 = E_i^0\psi_i^0, \tag{169}$$

where $\mathfrak{H}^0$ denotes, as usual, the operator obtained from $\mathcal{H}^0(q, p)$ by substituting $(h/2\pi i)(\partial/\partial q_r)$ for every p_r, according to the rules of §22. Let us suppose that we know how to solve the unperturbed problem completely, that is, that we know all the eigenvalues E_i^0 (which we assume to be discrete)[1] and the respective eigenfunctions ψ_i^0. Let us fix our attention on a definite state, the nth state for example, and let us determine the effect upon the latter due to the perturbing forces, that is, the modifications produced in E_n^0 and ψ_n^0. We shall assume in this section that the eigenvalue E_n^0 is not multiple. The case of multiple eigenvalues (*degeneracy*) requires separate treatment, which will be described in the following section. On the other hand, the other eigenvalues may be multiple, but in that case each of them will be counted as p coinciding eigenvalues (denoted by distinct indices), if it is multiple of order p.

The perturbing forces will be represented by a term $\mathcal{L}(q, p)$ added to the Hamiltonian, which will now be $\mathcal{H} = \mathcal{H}^0 + \mathcal{L}$. Hence the corresponding operator will become $\mathfrak{H}^0 + \mathfrak{L}$, where $\mathfrak{L}$ is obtained from $\mathcal{L}(q, p)$ by the usual substitution (including the eventual symmetrization of the products $p_r^m q)_r^n$, and is Hermitian. If, for instance, the perturbing forces are derivable from a potential $u(q)$, we shall have (recalling that $\mathcal{H}^0 = T + U$)

$$\mathfrak{L} = \mathcal{L} = u(q).$$

---

[1] If the eigenvalues are partly discrete and partly continuous, and if the eigenvalue which we are considering is discrete, the formulas of this and of the following section will still hold, provided we replace certain summations by integrals. If, however, the eigenvalue considered belongs to the continuous spectrum, a somewhat different procedure is required (see, for instance, No. 14 of the Bibliography, page 157).

In more general cases, as with magnetic forces, $\mathcal{L}$ will depend on the p in addition to the q, and therefore $\mathfrak{L}$ will also contain derivatives.

We shall suppose in every case that $\mathcal{L}$ does not contain t explicitly; this supposition may be expressed by saying that the perturbation is "time-independent." The opposite case will be discussed in §41.

The Schrödinger equation of the perturbed system for the nth stationary state will be (where $E_n = E_n^0 + \epsilon$ and $\psi_n = \psi_n^0 + \chi$ are called, respectively, the perturbed eigenvalue and the perturbed eigenfunction)[2]

$$(\mathfrak{H}^0 + \mathfrak{L})\psi_n = E_n\psi_n. \tag{170}$$

The function χ, which represents the effect of the perturbation upon ψ_n, may be expanded in a series in terms of the unperturbed eigenfunctions (which, as we know, form a complete orthogonal system) and may therefore be written

$$\psi_n = \psi_n^0 + \sum_{s=1}^{\infty} a_s\psi_s^0, \tag{171}$$

where the a_s are a set of constant coefficients which define the effect of the perturbation upon ψ_n^0. Substituting this expression into (170) and taking (169) into account (for $i = n$), we have

$$(-\epsilon + \mathfrak{L})\psi_n^0 + \sum_s a_s(E_s^0 - E_n^0 - \epsilon + \mathfrak{L})\psi_s^0 = 0.$$

Multiplying both terms by ψ_n^{0*} (from the left) and integrating over all q-space by making use of the conditions of orthogonality and normalization of the ψ, we indicate, as usual, by L_{rs} the elements of the matrix which represents the operator $\mathfrak{L}$ in the $\mathcal{H}^0$-representation (*perturbation matrix*), as follows:

$$L_{rs} = \int \psi_r^{0*}\mathfrak{L}\psi_s^0 \, dS. \tag{172}$$

We then get

$$\epsilon(1 + a_n) = L_{nn} + \sum_s a_s L_{ns}. \tag{173}$$

---

[2] We should write ϵ_n, χ_n, and later on, a_{rn}, because for every value of n there exists an ϵ, a χ, and a system of coefficients a. For the sake of simplicity we shall omit the index n, which may be considered fixed during the whole development. This practice will be followed frequently in the following pages.

Similarly, multiplying by $\psi_\sigma{}^{0*}$ (we indicate by σ an index taking on all positive integral values except n) and integrating, we have

$$a_\sigma(E_n^0 - E_\sigma{}^0 + \epsilon) = L_{\sigma n} + \sum_s a_s L_{\sigma s} \qquad (\sigma \neq n). \quad (174)$$

All the formulas derived thus far hold rigorously, no matter what the order of magnitude of $\mathfrak{L}$. Making use of the circumstance that the effect of the perturbation is small, let us consider the L_{rs}, ϵ, and the a_s as small quantities of the first order.[3] If we then neglect in (173) the quantities of the second order, that expression yields for ϵ a value in first approximation which we shall denote by ϵ':

$$\epsilon' = L_{nn}. \quad (175)$$

Therefore: *To a first approximation, the perturbation of the nth eigenvalue is given by the nth diagonal term of the perturbation matrix, or also, if desired* (see §27) *by the average value of $\mathfrak{L}$ calculated for the nth stationary state of the unperturbed system.* This result is entirely analogous to a known theorem of classical mechanics, according to which the correction to be applied to the energy of a system due to the effect of a perturbing force is equal to the average of the potential of this force, taken over the unperturbed motion.

It is to be noted that in order to calculate (in first approximation) the perturbed eigenvalues, it is not necessary to know the perturbed eigenfunctions ψ_n. Often the latter are not of interest and can be dispensed with. If we want to know them to a first approximation, we must solve (174) for the a_σ which, if the terms of second order are neglected and the first approximation of a_σ is denoted by a_σ', yields

$$a_\sigma' = \frac{L_{\sigma n}}{E_n^0 - E_\sigma{}^0} \qquad (\sigma \neq n). \quad (176)$$

In order to obtain all the terms of the expansion (171), we still have to know a_n. This is determined by imposing the condition upon ψ_n that it be normalized. Neglecting quantities of the second order and denoting, as usual, the first approximation by a prime,

---

[3] More precisely, we suppose all the $L_{n\sigma}$ to be small of first order compared with the differences $E_n{}^0 - E_\sigma{}^0$. It follows that ϵ and the a are also small of the first order (compared to $E_n{}^0$ and 1, respectively).

we then obtain

$$\int \left[\psi_n^{0*} \sum_{s=1}^{\infty} a_s' \psi_s^0 + \psi_n^0 \sum_{s=1}^{\infty} a_s'^* \psi_s^{0*}\right] dS = 0;$$

that is,

$$a_n' + a_n'^* = 0,$$

which means that a_n' must be pure imaginary but otherwise arbitrary (provided it is small of first order). If we take it to be zero,[4] we may write (171) as

$$\psi_n' = \psi_n^0 + \sum_{\sigma \neq n} \frac{L_{\sigma n}}{E_n^0 - E_\sigma{}^0} \psi_\sigma{}^0. \tag{177}$$

As far as the intuitive aspect is concerned, it might be mentioned that the effect of the perturbation upon the nth stationary state is one of "admixing" to the eigenfunction ψ_n^0 each of the other $\psi_\sigma{}^0$, in an amount which is the larger, the closer the respective energy levels lie to the level considered, and the more appreciable the element $L_{\sigma n}$ of the perturbation matrix.

Let us now proceed to the calculation of the eigenvalues in second approximation. In order to obtain ϵ from (173) in second approximation, it is necessary to insert the values of the a; but since the latter appear multiplied by quantities of the first order, it will be sufficient to introduce the values of the first approximation, namely, (176) and $a_n' = 0$. Denoting by ϵ'' the terms of second order of ϵ, we then obtain

$$\epsilon' + \epsilon'' = L_{nn} + \sum_{\sigma} \frac{|L_{\sigma n}|^2}{E_n^0 - E_\sigma{}^0} \qquad (\sigma \neq n). \tag{178}$$

Continuing in an analogous manner, we would calculate the eigenfunctions of the second approximation, and by means of the latter, the third approximation ($\epsilon' + \epsilon'' + \epsilon'''$) of ϵ, and so forth. These formulas are omitted here. They are rarely applied, since the first or second approximation for the eigenvalues and the first approximation for the eigenfunctions usually suffice.

---

[4] The arbitrariness of a_n reflects the arbitrary nature of the "phase" of $\psi_n{}^0$ and has no practical consequences. In fact, putting $a_n = i\delta_n$ (with δ_n real and small of first order), we obtain to a first approximation, $1 + a_n = e^{i\delta_n}$; and since in (171) $\psi_n{}^0$ appears just multiplied by $(1 + a_n)$, it suffices to substitute $\psi_n{}^0 e^{i\delta_n}$ for $\psi_n{}^0$ to get back to the case of $a_n = 0$.

39. Perturbation of stationary states (degenerate and quasi-degenerate case). In order for the approximations developed in the last section to be valid, it is necessary, as we have pointed out, that we have

$$|L_{\sigma n}| \ll |E_n^0 - E_\sigma{}^0|. \tag{179}$$

This necessity is apparent from formula (176), which shows that if for some σ this condition is not satisfied, the corresponding a_σ' is no longer a small quantity of the first order. The condition (179) will no longer be satisfied if the displacement of the nth level produced by the perturbation is of the same order of magnitude as the distance from the level to the neighboring levels. Therefore, if the level in question is rather close to some other level, even a slight perturbation will no longer permit the preceding arguments to be applied. As a limit of this situation, we may consider the case in which the level E_n is multiple (degeneracy), since we may imagine that we approach this condition gradually upon letting one or more of the other eigenvalues E_σ approach E_n until they coincide. Hence in this section we shall investigate how the perturbation method is to be modified in the case where there is a group of eigenvalues (which we shall call $E_1, E_2, \ldots E_p$) lying very close to one another, in such a way as to form a rather dense group called "multiplet" (*quasi-degeneracy*) or actually coinciding so as to form a multiple eigenvalue (*degeneracy*). Degeneracy will be considered as a limiting case of quasi-degeneracy. It is to be noted that these circumstances, which mathematically seem to be exceptional, are actually realized in the major part of problems of physical interest. In particular, energy levels are as a rule multiple in all problems with spherical or cylindrical symmetry (as we have seen, for instance, in the case of hydrogen). If we then take into account the corrections for spin and relativity which will be introduced in what follows, some of the coincident energy levels separate slightly, giving rise to the "fine-structure," or to multiplets. That is, we pass from complete degeneracy to the quasi-degenerate case. (Each component of the multiplet, however, may in turn present complete degeneracy.)

In this whole section let us adopt the convention that a letter $i, j, k, l, \ldots$ indicates any index (taking on the values $1, 2, \ldots p$) which serves to distinguish between the various components of the multiplet, and that the Greek letters $\rho, \sigma, \ldots$ will serve as indices (assuming all integral positive values except $1, 2, \ldots p$) referring to states which are not part of the multiplet. Let us then indicate (in conformity with the notation adopted above) by E_i^0 and $E_\sigma{}^0$ the unperturbed eigenvalues,[5] by E_i and E_σ the perturbed eigenvalues, and by ψ_i^0, $\psi_\sigma{}^0$, and ψ_i, ψ_σ, respectively, the eigenfunctions corresponding to them. It is then convenient to introduce the *average* energy level of the multiplet, that is, the quantity

$$E_0 = \frac{\sum\limits_i E_i^0}{p}, \tag{180}$$

and to put

$$E_i^0 = E_0 + \epsilon_i^0, \qquad E_i = E_0 + \epsilon_i. \tag{181}$$

In this way the quantities ϵ turn out to be small (of the first order) compared with the E. It is apparent that in the case of complete degeneracy, the ϵ_i^0 are all zero.

After these preliminaries, we write the Schrödinger equation for the unperturbed states as follows, denoting as before, the unperturbed Hamiltonian operator by $\mathfrak{H}^0$:

$$\mathfrak{H}^0\psi_i^0 = E_i^0\psi_i^0, \qquad \mathfrak{H}^0\psi_\sigma{}^0 = E_\sigma{}^0\psi_\sigma{}^0. \tag{182}$$

For the ith perturbed state we shall have instead

$$(\mathfrak{H}^0 + \mathfrak{L})\psi_i = E_i\psi_i. \tag{183}$$

Now it is to be noted that, in contrast to what occurs in the previous case, the perturbed eigenfunction ψ_i generally does not lie close to the unperturbed eigenfunction ψ_i^0, but differs from it (and from the other ψ_k^0) by terms which may not be considered small. This condition may be foreseen intuitively from the last section, since the closeness of other energy levels to the level E_i gives rise, in the summation of (177), to terms which are no longer small with respect to ψ_n^0. In other words, the various eigenfunctions of the multiplet "*mix*," without any one of them being predominant over the others. Therefore, if we expand ψ_i in a series by means of the

---

[5] The levels $E_\sigma{}^0$ may also be grouped, either entirely or partially, into multiplets or multiple levels. This fact does not change anything in the following formulas, provided that to every energy level of multiplicity n we make correspond n values of the index σ, as if dealing with distinct levels.

orthogonal functions $\psi_j^0, \psi_\sigma{}^0$ (which together form a complete system), we have

$$\psi_i = \sum_j c_{ij}\psi_j^0 + \sum_\sigma a_{i\sigma}\psi_\sigma{}^0. \tag{184}$$

We may consider the a as small of first order with respect to unity, and the c as of the order of magnitude 1 in general. Substituting into (183) and using (182), we obtain

$$\sum_j c_{ij}(E_j^0 - E_i + \mathfrak{L})\psi_j^0 + \sum_\sigma a_{i\sigma}(E_\sigma{}^0 - E_i + \mathfrak{L})\psi_\sigma = 0$$

or else

$$\sum_j c_{ij}(\epsilon_j^0 - \epsilon_i + \mathfrak{L})\psi_j^0 + \sum_\sigma a_{i\sigma}(\epsilon_\sigma{}^0 - \epsilon_i + \mathfrak{L})\psi_\sigma = 0.$$

Multiplying by ψ_k^{0*} and integrating over the whole range of the coordinates, we obtain, by making use of the orthogonality and normalization of the ψ^0, and introducing (172),

$$\sum_j c_{ij}[(\epsilon_j^0 - \epsilon_i)\delta_{jk} + L_{kj}] + \sum_\sigma a_{i\sigma}L_{k\sigma} = 0.$$

Since the L are small of the first order, like the a, the second summation will be negligible in first approximation. If we insert for the c_{ij} and ϵ_i the values of first approximation, which we shall call c_{ij}^0 and ϵ', we are left with

$$\sum_j c_{ij}^0[(\epsilon_j^0 - \epsilon_i')\delta_{jk} + L_{kj}] = 0. \tag{185}$$

Having fixed the value of i, and giving to k the p values $1, 2, \ldots p$, we obtain from this formula a system of p linear homogeneous equations in the p unknowns $c_{i1}^0, c_{i2}^0, \ldots c_{ip}^0$. In the coefficients of this system there occurs, besides the known quantities L_{kj}, ϵ_j^0, the still unknown quantity ϵ_j'.

Now for the system to have solutions which are not all zero, the determinant of the coefficients has to vanish; that is, we must have

$$\begin{vmatrix} L_{11} + \epsilon_1^0 - \epsilon_i' & L_{12} & L_{13} & \cdots \\ L_{21} & L_{22} + \epsilon_2^0 - \epsilon_i' & L_{23} & \cdots \\ L_{31} & L_{32} & L_{33} + \epsilon_3^0 - \epsilon_i' & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{vmatrix} = 0. \tag{186}$$

From this equation of degree p we can get ϵ_i'. This is of the form called "secular equation" (see §12); and since $L_{jk}^* = L_{kj}$, its p roots are all real (and for now we assume them to be distinct). It is to be noted that the p systems of equations which are obtained from (185) by giving to i the p values it may assume, all lead to the same secular equation, so that the p roots of equations (185) may be taken as the values of $\epsilon_1', \epsilon_2', \ldots \epsilon_p'$. In order to establish to which root each of the indices 1, 2, . . . p corresponds, it suffices to observe that if all the L_{ik} tend to zero, the last equation approaches

$$(\epsilon_1^0 - \epsilon_i')(\epsilon_2^0 - \epsilon_i') \cdots = 0,$$

and hence one of the roots tends toward ϵ_1^0, one toward ϵ_2^0, and so on. The correlation is made with the aid of this criterion.

Thus, from a solution of the secular equation (186), we have, to a first approximation, the perturbations of the eigenvalues of the multiplet.

From each of the systems (185) we then obtain the values $c_{i1}^0, c_{i2}^0, \ldots c_{ip}^0$ to within a constant factor. This factor is determined by imposing the condition

$$\sum_j |c_{ij}^0|^2 = 1. \tag{187}$$

With this the c_{ij}^0 will have the property[6]

$$\sum_j c_{ij}^{0*} c_{kj}^0 = \delta_{ik}, \tag{188}$$

which may be expressed by saying that they form the coefficients of an "orthogonal transformation."

If we then put

$$c_{ij} = c_{ij}^0 + c_{ij}',$$

where the first term is of the order unity and the second term is a small first-order correction, we may write (184)

$$\psi_i = \sum_j c_{ij}^0 \psi_j^0 + \sum_j c_{ij}' \psi_j^0 + \sum_\sigma a_{i\sigma} \psi_\sigma{}^0,$$

or also

$$\psi_i = \bar{\psi}_i + \sum_j c_{ij}' \psi_j^0 + \sum_\sigma a_{i\sigma} \psi_\sigma{}^0, \tag{189}$$

---

[6] Indeed, it may be shown that by using (185) and the relation $L_{ik} = L_{ki}{}^*$, we have $\sum_j c_{ij}{}^{0*} c_{kj}{}^0 = 0$ for $i \neq k$.

where we have put

$$\bar{\psi}_i = \sum_j c_{ij}^0 \psi_j^0. \tag{190}$$

In (189), $\bar{\psi}_i$ represents the principal term. As may be seen, the unperturbed eigenfunction is approximately (to within terms of the first order) equal, not to ψ_i^0, but to $\bar{\psi}_i$. The $\bar{\psi}_i$ may be called the eigenfunctions of the *zero-order* approximation.

The limiting case of complete degeneracy deserves separate consideration. In this case we deal, not with a multiplet, but actually with a level of multiplicity p. Here (185) becomes, since the ϵ_j^0 are zero,

$$\sum_j c_{ij}^0 [L_{kj} - \epsilon_i' \delta_{kj}] = 0, \tag{185'}$$

and the secular equation to be solved becomes

$$\begin{vmatrix} L_{11} - \epsilon_i' & L_{12} & L_{13} & \cdots \\ L_{21} & L_{22} - \epsilon_i' & L_{23} & \cdots \\ L_{31} & L_{32} & L_{33} - \epsilon_i' & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{vmatrix} = 0. \tag{200}$$

To its p real roots we may assign the indices 1, 2, . . . p in any order. The effect of the perturbation is to "split" the energy level E_0 into a group of p neighboring levels, given by $E_0 + \epsilon_1'$, $E_0 + \epsilon_2'$, . . . , to a first approximation. To the level E_0 there corresponded p stationary states, different as far as the ψ are concerned, but identical as to energy. By the effect of the perturbation each of these states acquires a slightly different energy. We may say that the perturbation *removes* the degeneracy.[7]

The following consideration will furnish an intuitive picture of the fact that the perturbed ψ_i are not, in general, approximately equal to the ψ_i^0, but rather to certain of their linear combinations $\bar{\psi}_i$. From §6 of Part II, we know that to a multiple eigenvalue (of order p) we may attribute an infinite number of systems of p (orthonormal) eigenfunctions, systems which may all be obtained

---

[7] If the secular equation had two or more coincident roots, the degeneracy would be only partially removed; that is, the level E_0 of multiplicity p would split into a group of levels, some of which would still be multiple (of order $< p$) in spite of the perturbation (at least to a first approximation). This case occurs frequently.

from any one of these systems by an orthogonal linear transformation. The ψ_i^0 represent any one of these systems (selected at random). Now we think of the perturbed state, which is not degenerate (hence the ψ_i represent a well-defined system of eigenfunctions). If the perturbation is permitted to go to zero, these ψ_i will tend toward a well-defined system of unperturbed eigenfunctions $\bar{\psi}_i$, and there is no reason why this should just be the system of the ψ_i^0. In general, it will be related to the latter by a certain linear orthogonal transformation. The systems of equations (185) merely serve to find the coefficients c_{ij}^0 of that transformation, and hence to determine, by means of (190), the $\bar{\psi}_i$, which by virtue of (188) will be normalized and orthogonal to each other (but not to the $\psi_\sigma{}^0$).

Let us now once again take up the general case, and let us try to find (to a first approximation) the perturbed eigenfunctions given by (189). It is convenient to transform this formula further by also expressing the second term in terms of the $\bar{\psi}$ instead of the ψ^0. Thereupon the formula becomes

$$\psi_i = \bar{\psi}_i + \sum_l \eta_{il}\bar{\psi}_l + \sum_\sigma a_{i\sigma}\psi_\sigma{}^0, \tag{201}$$

where the coefficients η_{il} (small of the first order) are related to the c_{ij}' by the following linear relations, which are found immediately by using (190):

$$c_{ij}' = \sum_l c_{lj}^0\eta_{il}. \tag{202}$$

Let us now express the fact that ψ_i satisfies the Schrödinger equation (183), namely, that

$$(\mathfrak{H}^0 - E_i + \mathfrak{L})\left(\bar{\psi}_i + \sum_l \eta_{il}\bar{\psi}_l + \sum_\sigma a_{i\sigma}\psi_\sigma{}^0\right) = 0.$$

Upon developing this expression, we note that, because of (190) and the first equation of (182),

$$\mathfrak{H}^0\bar{\psi}_i = \sum_j c_{ij}^0 E_j^0\psi_j^0 = E_0\bar{\psi}_i + \sum_j \epsilon_j^0 c_{ij}^0\psi_j^0,$$

and similarly for $\mathfrak{H}^0\Psi_l$. Furthermore, we recall the second equation (182) and take (202) into account. We then obtain

$$\left.\begin{aligned} -\epsilon_i\Psi_i + \sum_j \epsilon_j^0 c_{ij}^0 \psi_j^0 + L\Psi_i - \epsilon_i \sum_l \eta_{il}\Psi_l + \sum_j \epsilon_j^0 c'_{ij}\psi_j^0 \\ + \sum_l \eta_{il} L\Psi_l + \sum_\sigma a_{i\sigma}(E_\sigma{}^0 - E_i)\psi_\sigma{}^0 + \sum_\sigma a_{i\sigma} L\psi_\sigma{}^0 = 0. \end{aligned}\right\} \qquad (203)$$

From this we may first of all obtain the a by multiplying the equation by $\psi_\rho{}^{0*}$ and integrating. Then, since $\psi_\rho{}^0$ is orthogonal to Ψ_i, to the ψ_j^0, to the Ψ_l, and to all the $\psi_\sigma{}^0$ for which $\sigma \neq \rho$, we obtain

$$\bar{L}_{\rho i} + \sum_l \eta_{il}\bar{L}_{\rho l} + a_{i\rho}(E_\rho{}^0 - E_i) + a_{i\rho}L_{\rho\sigma} = 0, \qquad (204)$$

where we have introduced the following notation, analogous to (172):

$$\bar{L}_{\rho l} = \int \psi_\rho{}^{0*}\mathfrak{L}\Psi_l \, dS. \qquad (205)$$

Equation (204) was obtained without approximations. Its first and third terms are small quantities of the first order, the others of the second order. Neglecting the latter and replacing E_i by E_i^0, we obtain for the $a_{i\rho}$ the value in first approximation,

$$a'_{i\rho} = \frac{\bar{L}_{\rho i}}{E_i^0 - E_\rho{}^0} \qquad (206)$$

To solve for the η and the second approximation of the ϵ, we shall proceed in analogous fashion, multiplying (203) by Ψ_k^* and integrating. First, though, we note that

$$\int \Psi_k^* \psi_j^0 \, dS = \sum_l c_{kl}^{0*} \int \psi_l^{0*}\psi_j^0 \, dS = c_{kj}^{0*},$$

and we put

$$\bar{L}_{ki} = \int \Psi_k^* \mathfrak{L}\Psi_i \, dS. \qquad (205')$$

Hence we obtain, from (203),

$$\left.\begin{aligned} -\epsilon_i\delta_{ik} + \sum_j \epsilon_j^0 c_{ij}^0 c_{kj}^{0*} + \bar{L}_{ki} - \epsilon_i\eta_{ik} + \sum_j \epsilon_j^0 c'_{ij} c_{kj}^{0*} \\ + \sum_l \eta_{il}\bar{L}_{kl} + \sum_\sigma a_{i\sigma}\bar{L}_{k\sigma} = 0. \end{aligned}\right\} \qquad (207)$$

Now we see that because of (190), equation (205′) may be written thus:

$$\bar{L}_{ki} = \int \sum_{j,l} c_{kl}^{0*}\psi_l^{0*} c_{ij}^0 \mathfrak{L}\psi_j^0 \, dS = \sum_{j,l} c_{kl}^{0*} c_{ij}^0 L_{lj};$$

and since (185) may be written (by changing the index k into l)

$$\sum_j c_{ij}^0 L_{lj} = c_{il}^0(\epsilon_i' - \epsilon_l^0),$$

we have

$$\bar{L}_{ki} = \sum_l c_{kl}^{0*} c_{il}^0(\epsilon_i' - \epsilon_l^0),$$

and by changing the summation index l into j and using (188):

$$\bar{L}_{ki} = \epsilon_i'\delta_{ik} - \sum_j \epsilon_j^0 c_{kj}^{0*} c_{ij}^0. \tag{208}$$

If we substitute this expression for the third term of (207), we see that the second term of the latter cancels against the summation of (208). If relations (208) and (202) are taken into account, the next to last term of (207) transforms into

$$\sum_l \eta_{il}\bar{L}_{kl} = \sum_l \eta_{il}\left(\epsilon_l'\delta_{lk} - \sum_j \epsilon_j^0 c_{kj}^{0*} c_{lj}^0\right)$$
$$= \epsilon_k'\eta_{ik} - \sum_j \epsilon_j^0 c_{kj}^{0*} \sum_l \eta_{il} c_{lj}^0 = \epsilon_k'\eta_{ik} - \sum_j \epsilon_j^0 c_{kj}^{0*} c_{ij}'.$$

Thus (207) finally reduces to

$$-(\epsilon_i - \epsilon_i')\delta_{ik} - (\epsilon_i - \epsilon_k')\eta_{ik} + \sum_\sigma a_{i\sigma}\bar{L}_{k\sigma} = 0. \tag{209}$$

From this we obtain, for $k \neq i$,

$$\eta_{ik} = \frac{1}{\epsilon_i - \epsilon_k'} \sum_\sigma a_{i\sigma}\bar{L}_{k\sigma},$$

or else, in first approximation, utilizing (206),

$$\eta_{ik}' = \frac{1}{\epsilon_i' - \epsilon_k'} \sum_\sigma \frac{\bar{L}_{\sigma i}\bar{L}_{k\sigma}}{E_i^0 - E_\sigma^0} \qquad (i \neq k). \tag{210}$$

In order to complete our knowledge of the perturbed eigenfunctions in first approximation, we still need the coefficients η_{ii}, which are determined (like a_{nn} in the preceding section) from the normalization condition for ψ_i, and again we find that they may be taken to be zero.

Let us proceed to the determination of the perturbed eigenvalues in second approximation, which is, fairly important in practice. These eigenvalues are obtained from (209) for $k = i$. In fact, if we designate by ϵ_i'' the terms of second order and neglect those of higher order, that is, if we let $\epsilon_i = \epsilon_i' + \epsilon_i''$, then (209) yields, for $k = i$,

$$-\epsilon_i'' - \epsilon_i''\eta_{ii} + \sum_\sigma a_{i\sigma}\bar{L}_{i\sigma} = 0;$$

or, if we neglect the second term (which was taken to be zero and is of the third order in any case) and use (206),

$$\epsilon_i'' = \sum_\sigma \frac{|\bar{L}_{i\sigma}|^2}{E_i^0 - E_\sigma{}^0}. \tag{211}$$

Hence the second approximation for the eigenvalues of a multiplet or of a multiple level is given by

$$E_i = E_0 + \epsilon_i' + \sum_\sigma \frac{|\bar{L}_{i\sigma}|^2}{E_i^0 - E_\sigma{}^0}. \tag{212}$$

40. Perturbation theory by the matrix method. The perturbation theory of stationary states developed in the previous sections may naturally be presented from the point of view of the matrix method as well, which, as we know, leads to results equivalent to the results of wave mechanics. We shall now show this presentation as an example, confining ourselves to the first approximation and to the nondegenerate case.

If we assume for our system of reference in Hilbert space the system defined by the ψ_n^0, that is, if we refer to the "$\mathfrak{H}$-representation" of §33, the unperturbed Hamiltonian operator $\mathfrak{H}^0$ is represented by the diagonal matrix

$$\mathfrak{H}^0 = \begin{vmatrix} E_1^0 & 0 & 0 & \dots \\ 0 & E_2^0 & 0 & \dots \\ 0 & 0 & E_3^0 & \dots \\ \dots & \dots & \dots & \dots \end{vmatrix}.$$

The perturbed Hamiltonian operator $\mathfrak{H} = \mathfrak{H}^0 + \mathfrak{L}$ will instead be represented, with respect to the same axes, by a nondiagonal matrix $\{\mathfrak{H}\}$, in which, however, the off-diagonal elements are small compared with the diagonal elements. We are to effect a change of axes of reference such that with respect to the new axes, this operator will be represented by a diagonal matrix $\{\mathfrak{H}\}'$, whose elements will give the perturbed energy levels E_n. For this purpose we recall from §7 that to a change of axes there corresponds a transformation matrix $\{\mathfrak{S}\}$ and that the transform $\{\mathfrak{H}\}'$ of the matrix $\{\mathfrak{H}\}$ is given by $\{\mathfrak{H}\}' = \{\mathfrak{S}\}^{-1}\{\mathfrak{H}\}\{\mathfrak{S}\}$; that is, we have

$$\{\mathfrak{S}\}\{\mathfrak{H}\}' = \{\mathfrak{H}\}\{\mathfrak{S}\}. \tag{213}$$

In the actual case, since the matrix $\{\mathfrak{H}\}$ is *quasi*-diagonal, the transformation matrix rendering it diagonal will be little different from the unit matrix $\{1\}$ (or it will define a very slight rotation of the reference axes), and hence we shall write it in the form

$$\{\mathfrak{S}\} = \{1\} + \{\mathfrak{a}\}, \tag{214}$$

where the elements a_{rs} of the matrix $\{\mathfrak{a}\}$ are small quantities of the first order, which we are to determine. Equation (213) then becomes, if we substitute $\{\mathfrak{H}\} = \{\mathfrak{H}^0\} + \{\mathfrak{L}\}$ and neglect the product of the second order $\{\mathfrak{L}\}\{\mathfrak{a}\}$, and therefore replace $\{\mathfrak{a}\}$ by its first approximation $\{\mathfrak{a}'\}$,

$$\{\mathfrak{H}\}' = \{\mathfrak{H}^0\} + \{\mathfrak{L}\} + \{\mathfrak{H}^0\}\{\mathfrak{a}'\} - \{\mathfrak{a}'\}\{\mathfrak{H}\}',$$

or, since we may replace $\{\mathfrak{H}\}'$ by $\{\mathfrak{H}^0\}$ in the last term, making an error of the second order, we have

$$\{\mathfrak{H}\}' = \{\mathfrak{H}^0\} + \{\mathfrak{L}\} + \{\mathfrak{H}^0\}\{\mathfrak{a}'\} - \{\mathfrak{a}'\}\{\mathfrak{H}^0\}. \tag{215}$$

If we recall that the elements of $\{\mathfrak{H}^0\}$ are of the form $H^0_{rs} = E^0_r\delta_{rs}$, whereas the elements of $\{\mathfrak{H}\}'$ must be of the form $H'_{rs} = E_r\delta_{rs}$, we see that the relation (215) may be translated into the following relation between the elements:

$$E_n\delta_{nm} = E^0_n\delta_{nm} + L_{nm} + (E^0_n - E^0_m)a'_{nm}. \tag{216}$$

From this equation, we may simultaneously obtain the perturbed eigenvalues, even without determining the a'_{nm}. In fact, for $m = n$ relation (216) becomes

$$E_n = E^0_n + L_{nn}, \tag{217}$$

and thus we find the result (175). On the other hand, for $m \neq n$ we obtain

$$a'_{nm} = \frac{-L_{nm}}{E_n^0 - E_m^0}. \tag{218}$$

The diagonal elements a_{nn} remain arbitrary (provided they are real), and may be taken to be zero.

Having thus determined the transformation matrix $\{\mathfrak{S}\}$, we recall that the versors ψ_n of the rotated axes are obtained from those of the original axes by means of the formula (see §7)

$$\psi_n = \sum_m S_{mn}\psi_m^0,$$

or, since $S_{mn} = \delta_{mn} + a_{mn}$,

$$\psi_n = \psi_n^0 + \sum_m a_{mn}\psi_m^0.$$

Since $a_{nn} = 0$, it will be convenient to use, instead of the index m, the summation index σ (which excludes the value n). With this substitution and with the first approximation values (218) introduced for the a, the formula becomes identical with (177). We note that the coefficients a of §38 (where we have neglected to write the index n) are identified with the matrix elements of $\{\mathfrak{a}\}$ of (214), and hence define the infinitesimal rotation which brings us from the axes in which $\{\mathfrak{H}^0\}$ is diagonal to the axes in which $\{\mathfrak{H}\}$ is diagonal. This is the interpretation which, in Hilbert space, must be given to the process of approximation developed in §38.

41. Perturbation of nonstationary states. Time-dependent perturbations. Let us now treat the perturbation problem in a more general manner, so as to include also the case of states resulting from the superposition of several stationary states, and also the case of time-dependent perturbations. The method which we shall employ is known in mathematics as the "method of the variation of constants."

As before, let us indicate by ψ_r^0 the eigenfunctions of the unperturbed system, which are of the form

$$\psi_r^0 = u_r^0(q)\, e^{-\frac{2\pi i}{h}E_r^0 t} \tag{219}$$

and satisfy the equation

$$\mathfrak{H}^0\psi_r^0 = -\frac{h}{2\pi i}\frac{\partial\psi_r^0}{\partial t}. \tag{220}$$

Let us now introduce a perturbation, possibly also depending on the time, for which the Hamiltonian becomes

$$\mathcal{H}^0(q, p) + \mathcal{L}(q, p, t).$$

Here, for the first time, we are confronted by a Hamiltonian depending upon the time. We shall postulate that for this case the time-dependent Schrödinger equation also holds, in the same form as used thus far. The perturbed eigenfunction ψ will therefore satisfy the equation

$$(\mathfrak{H}^0 + \mathfrak{L})\psi = -\frac{h}{2\pi i}\frac{\partial \psi}{\partial t}. \tag{220′}$$

This ψ may be expanded in terms of the orthogonal functions ψ_r^0. Then we have

$$\psi = \sum_r c_r \psi_r^0, \tag{221}$$

where the coefficients c_r will in general be functions of t. Substituting this expansion into (220′) (and indicating, as we shall often do, the derivative with respect to time by a dot) we obtain, taking (220) into account,

$$\sum_r c_r \mathfrak{L} \psi_r^0 = -\frac{h}{2\pi i} \sum_r \dot{c}_r \psi_r^0.$$

From this the $\dot{c}$ may be obtained by multiplying both sides by ψ_s^{0*} and integrating over all q-space. We thus obtain

$$\dot{c}_s = -\frac{2\pi i}{h} \sum_r c_r L_{sr}, \tag{222}$$

where, as in the preceding sections, the quantities

$$L_{sr} = \int \psi_s^{0*} \mathfrak{L} \psi_r^0 \, dS \tag{223}$$

represent the elements of the "perturbation matrix" and may also be written, by virtue of (219),

$$L_{sr} = \lambda_{sr} e^{\frac{2\pi i}{h}(E_s^0 - E_r^0)t}, \tag{224}$$

with

$$\lambda_{sr} = \int u_s^{0*} \mathfrak{L} u_r^0 \, dS. \tag{225}$$

Let us suppose that for $t = 0$ the state of the system is represented by a certain $\psi(0)$ considered known, which when developed in a series in terms of the u_r^0 is

$$\psi(0) = \sum_r c_r^0 u_r^0. \tag{226}$$

Comparing with (221) and noting that the u_r^0 are the values of the ψ_r^0 for $t = 0$, we see that the coefficients c_r^0 of the expansion (226) (which are considered to be known) represent the initial values of the c_r. These initial values are to be associated with the differential equations (222) in order to obtain the c_r as functions of t, or the evolution of the state of the system. In the absence of perturbations, the c_r would evidently maintain their initial values, and hence the state of the (unperturbed) system at time t would be given by

$$\psi^0(t) = \sum_r c_r^0 \psi_r^0. \tag{227}$$

All these formulas are rigorous no matter what the nature of the perturbation. Let us now suppose that the perturbed state at time t differs little from the unperturbed state, or that $\psi(t)$ differs little from $\psi^0(t)$, that is, only by terms of the first order (hence this approximation will be valid for a time which is the longer, starting from $t = 0$, the weaker the perturbation). Then the $c_r(t)$ differ from the c_r^0 by terms of the first order, so that in the right-hand member of (222) we may replace $c_r(t)$ by c_r^0, introducing an error of the second order. We then obtain for the $\dot{c}_s$ the expressions

$$\dot{c}_s' = -\frac{2\pi i}{h} \sum_r c_r^0 L_{sr}$$

where the prime indicates that we are dealing with a first approximation. Integrating from 0 to t, we obtain the first-approximation values of the c_s:

$$c_s'(t) = c_s^0 - \frac{2\pi i}{h} \sum_r c_r^0 \int_0^t L_{sr}\, dt. \tag{228}$$

These, when inserted into (221), give the first approximation for the perturbed state.

Then, substituting (228) into the right hand side of (222) and integrating between 0 and t, we can easily obtain the second-order approximation, and similarly for successive ones; here, however, we shall confine ourselves to the first.

42. Transition probability. Let us now suppose that the perturbation lasts only for a certain interval of time, from 0 to t_1, while for $t < 0$ and $t > t_1$ we have $\mathcal{L} = 0$. We assume further that prior to the instant $t = 0$ the system is in the stationary state ψ_n^0, of energy E_n^0; that is, we assume that in (226) we have

$$c_n^0 = 1, \qquad c_\rho{}^0 = 0 \qquad (\text{for } \rho \neq n). \quad (229)$$

From the instant when the perturbation begins to act, this state is no longer stationary, and the ψ of the system for the time from 0 to t_1 may be written in the form (221), where the c_r, with equations (228) and the initial values (229) taken into consideration, are given in first approximation by

$$c_n'(t) = 1 - \frac{2\pi i}{h}\int_0^t L_{nn}\,dt,$$

$$c_\rho'(t) = -\frac{2\pi i}{h}\int_0^t L_{\rho n}\,dt \qquad (\rho \neq n).$$

From the instant t_1 on, the coefficients c_n, c_ρ become constants; but they will have, rather than the values (229), the values obtained from the preceding formulas in which t was replaced by t_1. We shall call them c_n^1, $c_\rho{}^1$; that is, we shall put

$$c_n^1 = 1 - \frac{2\pi i}{h}\int_0^{t_1} L_{nn}\,dt, \quad (230)$$

$$c_\rho{}^1 = -\frac{2\pi i}{h}\int_0^{t_1} L_{\rho n}\,dt \qquad (\rho \neq n). \quad (230')$$

This means that if, after the instant t_1, we were to make a new determination of the state, there would exist a certain probability, given in first approximation by $|c_\rho{}^1|^2$, of finding the system in the state $\psi_\rho{}^0$ rather than in the initial state ψ_n^0. If the observation of the state gives this result, we shall say that a *transition* from the nth to the ρth state has occurred. The effect of the perturbation is therefore one of inducing a certain *transition probability* between one state (stationary in the absence of perturbation) and all the others; and the amplitude of this probability for the transition $n \rightarrow \rho$ is determined, as is shown by (230′), by the element $L_{\rho n}$ of the perturbation matrix. The probability of the inverse transition $\rho \rightarrow n$ is evidently the same, since $L_{n\rho} = L_{\rho n}{}^*$.

Note the difference between this result of quantum mechanics and the point of view of the old Bohr-Sommerfeld theory, which postulated that the system remain in the nth stationary state until it "jumped" into the ρth state by a sudden process. Quantum mechanics assumes, in addition to the pure stationary states, the existence of states resulting from the superposition of different stationary states, and views the effect of the perturbation as a continuous process by which upon the nth pure state there are gradually superimposed varying "amounts" of other states. The sudden transition into one of the latter will then happen only at the instant of observation, by the effect of the inevitable action of the instruments of observation on the observed system.

43. Perturbation of the sinusoidal type. Resonance.[8] Let us apply the results of the preceding section to the case in which the perturbing force is a sinusoidal function of time, of frequency ν. This case occurs, for instance, when an atom is exposed to monochromatic radiation.

Let us therefore suppose that the perturbing term of the Hamiltonian is of the form

$$\mathfrak{L} = A(q, p) \cos 2\pi\nu t.$$

The elements of the perturbation matrix are [see formulas (224) and (225)]

$$L_{sr} = A_{sr}\, e^{\frac{2\pi i}{h}(E_s{}^0 - E_r{}^0)t} \cos 2\pi\nu t,$$

with A_{sr} independent of time, or also, setting

$$\frac{E_s^0 - E_r^0}{h} = \nu_{sr},$$

$$L_{sr} = \tfrac{1}{2}A_{sr}[e^{2\pi i(\nu_{sr}+\nu)t} + e^{2\pi i(\nu_{sr}-\nu)t}]. \tag{231}$$

The transition-probability amplitude from the state n to the state ρ, after a time t_1 of perturbing action, is given by (230′), which

---

[8] The term "resonance" is used here in the classical sense. In quantum mechanics it also has another meaning, which will be illustrated in Chapter 15.

in that case yields

$$c_p{}^1 = -\frac{A_{pn}}{2h}\left[\frac{e^{2\pi i(\nu_{pn}+\nu)t_1}-1}{\nu_{pn}+\nu}+\frac{e^{2\pi i(\nu_{pn}-\nu)t_1}-1}{\nu_{pn}-\nu}\right]$$
$$= -\frac{iA_{pn}}{h}\left[e^{i\pi(\nu_{pn}+\nu)t_1}\frac{\sin\pi(\nu_{pn}+\nu)t_1}{\nu_{pn}+\nu}+e^{i\pi(\nu_{pn}-\nu)t_1}\frac{\sin\pi(\nu_{pn}-\nu)t_1}{\nu_{pn}-\nu}\right]. \tag{232}$$

Now it is known that the expression $\left|\frac{\sin\pi\alpha t_1}{\alpha}\right|$, considered as a function of α, has an absolute maximum equal to πt_1, for $\alpha = 0$, whereas the other maxima are considerably lower (the qualitative behavior is the same as that of the curve in Fig. 20). Therefore the two terms in the square brackets can take on appreciable values only for $\nu = -\nu_{pn}$ (for the first term) and for $\nu = \nu_{pn}$ (for the second term). The first of these cases may be realized if $E_p{}^0 < E_n^0$, the second case if $E_p{}^0 > E_n^0$. In both cases we have, neglecting the other term,

$$c_p{}^1 = -\frac{i\pi A_{pn}}{h}t_1.$$

This means that the transition probability from the state n to the state p is appreciable only if the energy difference between the two states, $|E_p{}^0 - E_n^0|$, is very close to the value $h\nu$, or if $|\nu_{pn}|$ is very close to ν. In practice, the maximum is so sharp that the transition must be considered possible *only* if $|\nu_{pn}| = \nu$ (condition for *resonance*). This transition, if $E_p{}^0 > E_n^0$, requires the absorption of energy in the form of radiation. From this it follows that *the atom in the state n can absorb only radiation of frequency $\nu = (1/h)(E_p{}^0 - E_n^0)$, and absorbs it in quanta $h\nu$.* Thus Bohr's postulate concerning the relation between energy and frequency is shown to be (for the case of absorption) a consequence of wave mechanics.

If $E_p{}^0 < E_n^0$, the transition $n \rightarrow p$ takes place with the emission of radiation. The theory discussed here does not furnish the elements for calculating the frequency of this radiation; but the complete theory of the interaction of a radiation field and an atom due to Dirac (mentioned in §32 of Part II), makes it possible to show the validity of the Bohr postulate for this latter case as well. In addition, it shows the possibility of "spontaneous" transitions (with emission) from the state n to a lower-lying one, whereas the transitions dealt with here are induced by external radiation.

CHAPTER 14

Relativity and Spin

44. General considerations. Quantum mechanics as developed in the preceding chapters was constructed by starting from the classical (nonrelativistic) mechanics of a material point; as we know, it reduces to the classical case whenever a wave packet may be considered to be a point. However, it is known that classical mechanics, which holds for velocities which are small compared with the velocity of light *c*, constitutes only a first approximation of relativistic mechanics, which is valid for motions with any velocity. Hence we must remember that quantum mechanics as developed so far is valid only under the same limitations, and that in order to obtain a more rigorous and general quantum mechanics we must start from the relativistic mechanics of a point particle, rather than from classical mechanics. The necessity for this refinement becomes evident when we consider that the results of Schrödinger wave mechanics are not invariant under a Lorentz transformation.

Another fact which was partly neglected in the preceding chapters is the existence of an intrinsic angular momentum (*spin*) and of a magnetic moment, both in the electron and in the proton, and presumably in other particles as well. This spin, as was pointed out in §25 of Part I and in §62 of Part II, is proved by many facts of spectroscopy and electromagnetism, and was first postulated by Uhlenbeck and Goudsmit under the name "hypothesis of the spinning electron."

At first, an attempt was made to deal separately with these two causes of inexactitude of quantum mechanics. Thus, extending the analogy to ordinary mechanics, many workers attempted a modification of the Schrödinger equation, in order to take the relativistic correction into account. This effort, which we shall discuss briefly in §46, was only partially successful. On the other hand, Pauli succeeded in introducing the spin hypothesis into (nonrelativistic) quantum mechanics, constructing a remarkable theory which will be presented in §45. But the most satisfactory solution of both these questions was found by Dirac, who showed that the two modifications—the one concerning relativity and the one concerning the spinning electron—are conceptually reduced to one and the same modification. This identity results because, when wave mechanics is given a suitable relativistic form, there follows the existence of the spin and of the magnetic moment, with their correct values and rules for spatial quantization, without the necessity of introducing them by an *ad hoc* hypothesis. From the Dirac theory we may then obtain the Pauli theory as a first nonrelativistic approximation. First, however, we shall deal with the Pauli theory, because it will furnish a simple example of certain new mathematical methods which are used on a much larger scale in the Dirac theory.[1]

45. Fundamentals of nonrelativistic spin theory (Pauli). We recall that the essential point in the hypothesis of Uhlenbeck and Goudsmit (from which are derived the spectroscopic consequences which have lent credence to the hypothesis) is that *the projection of the spin upon any direction* always has one of the two values $\pm h/4\pi$, and correspondingly, the projection of the magnetic moment upon the same direction has the value $\pm\mu_0$ (where μ_0 stands for the absolute value of the Bohr magneton). Let us therefore introduce three new observables σ_x, σ_y, σ_z, representing the projections of the spin upon the three axes, *measured in units of* $h/4\pi$; and in order to reflect the Uhlenbeck-Goudsmit hypothesis, we shall assume that each of these has only the eigenvalues ± 1. We call them "spin components" (implying "in units of $h/4\pi$"), and, according to common usage, we shall denote the operators corresponding to them by the same symbols. Furthermore, we shall introduce the components μ_x, μ_y, μ_z of the magnetic moment,

---

[1] Essentially, the Dirac theory deals with the case of a single electron. Until quite recently a rigorous and completely relativistic theory of a system of several electrons was lacking. However, during 1948 and 1949 great progress was made in that direction through the contributions of Tomonaga and others in Japan, and Schwinger, Feynman, Dyson, and others in the United States. Following is a list of a few of the most important papers on this subject: J. Schwinger, *Phys. Rev.* **74**, 1439 (1948); *ibid.*, **75**, 651 (1949); *ibid.*, **76**, 790 (1949); R. P. Feynman, *Phys. Rev.* **76**, 749 (1949); *ibid.*, **76**, 769 (1949); F. J. Dyson, *Phys. Rev.* **75**, 486 (1949); *ibid.*, **75**, 1736 (1949). Further references may be found in some of these papers.

as three observables (or three operators) related to the σ by the formulas

$$\mu_x = -\mu_0\sigma_x, \qquad \mu_y = -\mu_0\sigma_y, \qquad \mu_z = -\mu_0\sigma_z. \tag{233}$$

Let us now investigate the properties of the operators σ_x, σ_y, σ_z, defined in this manner. First we note that since they have only the two eigenvalues ± 1, their squares will have the unique eigenvalue $+1$, so that it will be legitimate to write:[2]

$$\sigma_x^2 = \sigma_y^2 = \sigma_z^2 = 1. \tag{234}$$

Now it is necessary to establish the commutation rules of the operators σ_x, σ_y, σ_z. Pauli did this by assuming that the components of the spin behave in this respect like the components of an ordinary angular momentum, which, as we have shown in §30, satisfy the commutation relations (125). Since these relations hold for angular momenta in ordinary units, we must substitute $(h/4\pi)\sigma_x$ for $\mathfrak{M}_x$, and so on. In this way we find the following commutation relations:

$$\left.\begin{aligned} \sigma_y\sigma_z - \sigma_z\sigma_y &= 2i\sigma_x, \\ \sigma_z\sigma_x - \sigma_x\sigma_z &= 2i\sigma_y, \\ \sigma_x\sigma_y - \sigma_y\sigma_x &= 2i\sigma_z. \end{aligned}\right\} \tag{235}$$

Let us now multiply the second relation by σ_z (from the right), and the third one by σ_y; upon adding them we get

$$\sigma_z\sigma_x\sigma_z - \sigma_y\sigma_x\sigma_y + \sigma_x(\sigma_y^2 - \sigma_z^2) = 2i(\sigma_y\sigma_z + \sigma_z\sigma_y).$$

Substituting into the first two terms the expression for σ_x obtained from the first relation (235), and recalling (234), we recognize that the whole left-hand side vanishes, and hence there remains

$$\sigma_y\sigma_z + \sigma_z\sigma_y = 0.$$

Two other relations analogous to this one may be obtained in the same way. Therefore the components of the spin *anticommute.*

---

[2] Incidentally, it follows that $\sigma_x^2 + \sigma_y^2 + \sigma_z^2 = 3$, and hence the *total spin* σ (in units of $h/4\pi$), defined by $\sigma^2 = \sigma_x^2 + \sigma_y^2 + \sigma_z^2$, turns out to be $\sqrt{3}$ and not 1 as is usually assumed in the vector model. This discrepancy is due to the inadequacy of this model, which was mentioned several times. In fact, we note that if in the theory of Uhlenbeck and Goudsmit we use the total spin instead of the projection of the spin on a given direction, we obtain formulas requiring slight modifications in order to agree with experience (see, for example, No. 27 of the Bibliography, page 198).

Upon using (235) again, we may write

$$\left.\begin{aligned} \sigma_y\sigma_z &= -\sigma_z\sigma_y = i\sigma_x \\ \sigma_z\sigma_x &= -\sigma_x\sigma_z = i\sigma_y \\ \sigma_x\sigma_y &= -\sigma_y\sigma_x = i\sigma_z. \end{aligned}\right\} \tag{236}$$

When we take the spin into consideration, a measurement of the coordinates x, y, z of the electron is no longer a "maximum observation," since it does not yet completely define the state of the electron. In order to make it complete, it is necessary to add an observation of the projection of the spin (or of the magnetic moment, which amounts to the same thing) along any direction, which is usually chosen so as to coincide with the z-axis. The determination of the state is therefore complete with an observation of σ_z (note that we may not add an observation of σ_x and of σ_y; the latter are incompatible with the former, because their respective operators do not commute). The state will therefore now be defined by a function $\psi(x, y, z, \sigma_z, t)$ rather than only by $\psi(x, y, z, t)$. However, since the newly introduced variable (called *spin variable*) can have only the two values ± 1, it is often convenient to consider it as an *index* rather than as an ordinary variable, that is, to write $\psi_1(x, y, z, t)$ and $\psi_2(x, y, z, t)$ instead of $\psi(x, y, z, +1, t)$ and $\psi(x, y, z, -1, t)$. The introduction of the spin variable is therefore equivalent to introducing two functions ψ_1, ψ_2 instead of only one ψ; this is the main feature of the Pauli theory (see §25 of Part II). The significance of these two functions is obviously as follows: $|\psi_1|^2 dx\, dy\, dz$ represents the probability of finding the electron in the element of volume defined by $x, y, z, x + dx, y + dy, z + dz$, and with spin $\sigma_z = 1$ (or else with $\mu_z = -\mu_0$), whereas $|\psi_2|^2 dx\, dy\, dz$ is the probability of finding it in the same element of volume, but with $\sigma_z = -1$ (or with $\mu_z = +\mu_0$). Using the terminology of the vector model, we would say that in the first case the spin is *parallel* to the z-axis, in the second case *antiparallel* to it. The privileged role given to the z-axis may of course just as well be given to the x- or y-axes, or to any other direction. It would then be necessary to use another pair of functions ψ.

The normalization condition is evidently

$$\iiint (|\psi_1|^2 + |\psi_2|^2)\, dx\, dy\, dz = 1. \tag{237}$$

It is sometimes convenient (in analogy to what was done in §6a) to consider the pair ψ_1, ψ_2 as a matrix with a single column and two rows, which is indicated symbolically by ψ, that is, to put

$$\psi = \begin{vmatrix} \psi_1 \\ \psi_2 \end{vmatrix}. \tag{238}$$

In correspondence with this representation, we shall find it convenient to represent the operators σ_x, σ_y, σ_z by means of matrices, noting that they have only two eigenvalues and will therefore be represented by matrices with only two rows and two columns, of the form

$$\begin{vmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{vmatrix},$$

where the k are constants which, in order to be Hermitian, must satisfy the condition $k_{21} = k_{12}^*$. These operators, when applied to (238), replace the functions ψ_1, ψ_2 by two of their linear combinations, following the scheme (in accordance with the rule for matrix multiplication)

$$\begin{vmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{vmatrix} \begin{vmatrix} \psi_1 \\ \psi_2 \end{vmatrix} = \begin{vmatrix} k_{11}\psi_1 + k_{12}\psi_2 \\ k_{21}\psi_1 + k_{22}\psi_2 \end{vmatrix}. \tag{239}$$

(Since in this process ψ_1 and ψ_2 behave as if they did not depend upon x, y, z, we shall say that such an operator "operates only upon the spin variable.") Let us proceed to the actual determination of these three matrices, which are denoted by the same symbols σ_x, σ_y, σ_z as the operators they represent. We observe first of all that ψ_1 and ψ_2, because of the meaning given them above, are simply the eigenfunctions of the operator σ_z, corresponding to the eigenvalues $+1$ and -1, respectively; that is, we must have

$$\sigma_z\psi_1 = \psi_1, \qquad \sigma_z\psi_2 = -\psi_2,$$

or else

$$\sigma_z \begin{vmatrix} \psi_1 \\ \psi_2 \end{vmatrix} = \begin{vmatrix} \psi_1 \\ -\psi_2 \end{vmatrix}. \tag{240}$$

From this we deduce, by virtue of (239),

$$\sigma_z = \begin{vmatrix} 1 & 0 \\ 0 & -1 \end{vmatrix}. \tag{241}$$

In order to determine the matrices σ_x and σ_y, let us first write them in the general Hermitian form

$$\sigma_x = \begin{vmatrix} a_1 & a_3 \\ a_3^* & a_2 \end{vmatrix}, \qquad \sigma_y = \begin{vmatrix} b_1 & b_3 \\ b_3^* & b_2 \end{vmatrix},$$

and let us require each of these to anticommute with σ_z. By (241), we have

$$\sigma_x\sigma_z = \begin{vmatrix} a_1 & -a_3 \\ a_3^* & -a_2 \end{vmatrix}, \qquad \sigma_z\sigma_x = \begin{vmatrix} a_1 & a_2 \\ -a_3^* & -a_2 \end{vmatrix},$$

and hence, since $\sigma_x\sigma_z + \sigma_z\sigma_x = 0$, we must have $a_1 = a_2 = 0$. We then have

$$\sigma_x^2 = \begin{vmatrix} a_3a_3^* & 0 \\ 0 & a_3a_3^* \end{vmatrix},$$

and in order for this to reduce to a unit matrix, we must have $a_3a_3^* = 1$, that is, $a_3 = e^{i\alpha}$, with α real (and arbitrary). If we proceed in analogous fashion for σ_y, it may be concluded that we must have

$$\sigma_x = \begin{vmatrix} 0 & e^{i\alpha} \\ e^{-i\alpha} & 0 \end{vmatrix}, \qquad \sigma_y = \begin{vmatrix} 0 & e^{i\beta} \\ e^{-i\beta} & 0 \end{vmatrix}.$$

Then, imposing the condition $\sigma_x\sigma_y + \sigma_y\sigma_x = 0$, we find between α and β the relation $\alpha - \beta = \pi/2 + n\pi$ (with n any integer). The arbitrariness of n which remains in one of the two constants α, β, has no physical consequences.[3] Taking $\alpha = 0$, $\beta = -\pi/2$, we finally obtain

$$\sigma_x = \begin{vmatrix} 0 & 1 \\ 1 & 0 \end{vmatrix}, \qquad \sigma_y = \begin{vmatrix} 0 & -i \\ i & 0 \end{vmatrix}. \tag{241'}$$

These, together with (241), give the desired expressions. It is apparent that σ_z is of the diagonal form, since because of the privileged role which we have assigned to the z-axis, the matrices are referred to the "σ_z-representation." If some other representation were adopted (and hence another meaning for ψ_1 and ψ_2), the three matrices would transform as was explained in §8.

By an obvious extension of the principles of §22, we may obtain the operator corresponding to any quantity relative to the spin by writing down the classical expression for this quantity as a function

---

[3] This condition corresponds to the arbitrariness of the constant θ in the argument of ψ, mentioned on page 155.

of the spin components and replacing them by their corresponding operators. For instance, the projection of the spin upon any direction n, with direction cosines α, β, γ, will be represented by the operator

$$\sigma_n = \alpha\sigma_x + \beta\sigma_y + \gamma\sigma_z, \tag{242}$$

or, in the σ_z-representation, by the matrix

$$\sigma_n = \begin{vmatrix} \gamma & \alpha - i\beta \\ \alpha + i\beta & -\gamma \end{vmatrix}, \tag{242'}$$

which may be immediately obtained by substituting (241) and (241′) into (242).

As an application, let us suppose that an observation of the spin with respect to a certain direction n, with direction cosines α, β, γ, has given the result $+1$, and that immediately afterward an observation of the spin is carried out with respect to the z-axis. What is the probability of finding a value $+1$, and what is the probability of finding -1? The state resulting from the first observation will be defined by an expression $\begin{vmatrix}\psi_1 \\ \psi_2\end{vmatrix}$ such that (see §22)

$$\sigma_n \begin{vmatrix}\psi_1 \\ \psi_2\end{vmatrix} = 1 \begin{vmatrix}\psi_1 \\ \psi_2\end{vmatrix},$$

which is equivalent, when (242′) is applied, to

$$\left.\begin{aligned} \gamma\psi_1 + (\alpha - i\beta)\psi_2 &= \psi_1, \\ (\alpha + i\beta)\psi_1 - \gamma\psi_2 &= \psi_2. \end{aligned}\right\}$$

These two homogeneous equations (whose determinant is zero by virtue of $\alpha^2 + \beta^2 + \gamma^2 = 1$) yield

$$\frac{\psi_1}{\psi_2} = \frac{\alpha - i\beta}{1 - \gamma},$$

and hence the ratio of the probabilities of the two results $+1$ and -1 is

$$\left|\frac{\psi_1}{\psi_2}\right|^2 = \frac{\alpha^2 + \beta^2}{(1 - \gamma)^2} = \frac{1 + \gamma}{1 - \gamma}.$$

If, in particular, the direction n were normal to the z-axis, the two probabilities would be equal.

It is to be noted that all these results cannot be interpreted by the simple vector model, according to which the second observation would certainly give the result $\sigma_z = \gamma$.

The Hamiltonian of an electron having a spin and situated in a magnetic field may be written down by analogy to its expression in classical mechanics, which in first approximation[4] may be obtained by simply adding to the Hamiltonian of an electron without spin (§31) the interaction energy between its magnetic moment and the magnetic field. That is, if we call m_0 the mass, $-e$ the charge,[5] **E** the electric field, **H** the magnetic field, V the scalar potential and **A** the vector potential, and $\boldsymbol{\mu}$ the magnetic moment, we shall obtain

$$\mathcal{H} = \frac{1}{2m_0} \sum_{r=1}^{3} \left(p_r + \frac{e}{c} A_r\right)^2 - eV - \boldsymbol{\mu} \cdot \mathbf{H}. \tag{243}$$

If p_r and μ_r are replaced by the corresponding operators $\frac{h}{2\pi i}\frac{\partial}{\partial x_r}$, $-\mu_0\sigma_r$, this expression transforms into the operator

$$\mathfrak{H} = \frac{1}{2m_0} \sum_{r=1}^{3} \left(\frac{h}{2\pi i}\frac{\partial}{\partial x_r} + \frac{e}{c} A_r\right)^2 - eV + \mu_0 \boldsymbol{\sigma} \cdot \mathbf{H} \tag{244}$$

where, extending the notation for the scalar product in an obvious manner, we have indicated by the symbol $\boldsymbol{\sigma} \cdot \mathbf{H}$ the operator (or matrix)

$$\boldsymbol{\sigma} \cdot \mathbf{H} = \sum_{r=1}^{3} \sigma_r H_r = \begin{vmatrix} H_z & H_x - iH_y \\ H_x + iH_y & -H_z \end{vmatrix}. \tag{245}$$

The Hamiltonian operator $\mathfrak{H}$ formed in this way then allows us to write the equation for ψ in the usual manner, namely,

$$\mathfrak{H}\psi = -\frac{h}{2\pi i}\frac{\partial \psi}{\partial t}, \tag{246}$$

and for stationary states

$$\mathfrak{H}\psi = E\psi. \tag{246'}$$

---

[4] The rigorous formula would also contain terms of the order v/c compared with the others, representing the action of the electric field upon the magnetic moment in motion.

[5] In this whole chapter, we shall indicate by e the charge of the electron in *absolute value*, and by m_0 its *rest mass*.

Equation (246), as well as (246′), may formally be considered to be an equation in $\psi(x, y, z, \sigma_z, t)$, or else, more explicitly, as a system of two equations in the two functions $\psi_k(x, y, z, t)$ (with $k = 1, 2$). For example, if we indicate by $\mathfrak{H}_0$ the part of the Hamiltonian (244) which does not operate on the spin, that is, if we write $\mathfrak{H} = \mathfrak{H}_0 + \mu_0 \boldsymbol{\sigma} \cdot \mathbf{H}$, we may write (246) more explicitly, by means of (245), as two equations:

$$\left.\begin{aligned} \mathfrak{H}_0\psi_1 + \mu_0 H_z\psi_1 + \mu_0(H_x - iH_y)\psi_2 &= -\frac{h}{2\pi i}\frac{\partial\psi_1}{\partial t}, \\ \mathfrak{H}_0\psi_2 + \mu_0(H_x + iH_y)\psi_1 - \mu_0 H_z\psi_2 &= -\frac{h}{2\pi i}\frac{\partial\psi_2}{\partial t}, \end{aligned}\right\}$$

and similarly for (246′). If the magnetic field is zero or negligible, each of the two ψ satisfies the ordinary Schrödinger equation. Hence, if we are dealing with a nondegenerate[6] stationary state, these two must differ only by a constant coefficient.

Separation of the spin variable. If the magnetic field is uniform, the part of the Hamiltonian depending on the spin will not contain the position coordinates, and it is then possible (for a stationary state) to perform the "separation" of the spin variable σ_z from x, y, z, that is, to write $\psi(x, y, z, \sigma_z, t)$ in the form $\Psi_n(x, y, z, t)\varphi_s(\sigma_z)$, where $\sigma_z = \pm 1$ and hence $\varphi_s(\sigma_z)$ represents, rather than a true function, the total of two constants $\varphi_s(1) = \alpha_s$, $\varphi_s(-1) = \beta_s$, or the matrix $\begin{vmatrix}\alpha_s \\ \beta_s\end{vmatrix}$. The index s is the spin quantum number which, as we shall see, may take on only two values, whereas n stands as usual for the group of three orbital quantum numbers.

Indeed (246′) will now split into the two expressions

$$\mathfrak{H}_0\Psi_n = E'_n\Psi_n \tag{247}$$

$$\mu_0\boldsymbol{\sigma} \cdot \mathbf{H}\varphi_s = E''_s\varphi_s \tag{248}$$

with $E'_n + E''_s = E$. The first of these is the ordinary Schrödinger equation. Hence E'_n represents one of the energy levels of ordinary wave mechanics, and Ψ_n is an eigenfunction which corresponds to it. The second equation (where E''_s represents the energy due to the action of the magnetic field upon the magnetic moment of the spin)

---

[6] The degeneracy due to the spin is understood to be excluded.

may be more explicitly written as two algebraic equations

$$\left.\begin{aligned}\mu_0 H_z \alpha_s + \mu_0(H_x - iH_y)\beta_s &= E_s''\alpha_s,\\ \mu_0(H_x + iH_y)\alpha_s - \mu_0 H_z\beta_s &= E_s''\beta_s,\end{aligned}\right\} \tag{249}$$

which are linear and homogeneous in α_s and β_s. Since the latter are not both zero, we must have

$$\begin{vmatrix}\mu_0 H_z - E_s'' & \mu_0(H_x - iH_y)\\ \mu_0(H_x + iH_y) & -\mu_0 H_z - E_s''\end{vmatrix} = 0,$$

that is, $E_s''^2 = \mu_0^2(H_x^2 + H_y^2 + H_z^2) = \mu_0^2 H^2$. From this we get the two values of E_s'', corresponding to the two values 1 and 2 of the index s:

$$E_1'' = \mu_0 H, \qquad E_2'' = -\mu_0 H. \tag{250}$$

This result justifies the success of modelistic spin theory, in which it was postulated that the spin could line up parallel or antiparallel to the field, the two values (250) of the magnetic energy corresponding to these two cases.

If we now assume the magnetic field to be directed along the z-axis, and if we solve the system (249) (determining the normalization constant such that $|\alpha_s|^2 + |\beta_s|^2 = 1$), we find

for E_1'': $\alpha_1 = 1, \quad \beta_1 = 0,$

for E_2'': $\alpha_2 = 0, \quad \beta_2 = 1.$

Hence we have, in the first case, $\psi_{n,1} = \begin{vmatrix}\Psi_n\\ 0\end{vmatrix}$, which means that the spin is certainly directed along the positive z-axis, and in the second case $\psi_{n,2} = \begin{vmatrix}0\\ \Psi_n\end{vmatrix}$, where the spin is directed with certainty in the opposite direction.

46. The relativistic extension of the Schrödinger equation. Before exploring the Dirac theory, we shall show what equation for ψ we should obtain if, applying the principle of §22, we started out from the relativistic expression for the Hamiltonian (rather than from the classical expression, as was done in §19), and if we transformed it into an operator by means of the usual substitutions (S), (S') of §19.

Let us consider a particle[7] of charge ϵ (in electrostatic units) and (rest) mass m_0, placed in an electric and magnetic field, derived from a scalar potential V and a vector potential **A**. The relativistic Hamiltonian is[8]

$$\mathcal{H} = c\sqrt{m_0^2c^2 + \sum_{k=1}^{3}\left(p_k - \frac{\epsilon}{c}A_k\right)^2} + \epsilon V, \tag{251}$$

where the momenta p_k are given by (see §31)

$$p_k = \frac{m_0 v_k}{\sqrt{1 - v^2/c^2}} + \frac{\epsilon}{c}A_k. \tag{252}$$

The energy integral is

$$\mathcal{H} - E = 0. \tag{253}$$

However, since the first term is irrational in the p_k, it is convenient, before performing the substitutions (S), (S'), to rationalize it by isolating the radical and squaring. Hence we start, instead of from (253), from the relation

$$\frac{1}{c^2}(\epsilon V - E)^2 - \sum_k\left(p_k - \frac{\epsilon}{c}A_k\right)^2 - m_0^2c^2 = 0. \tag{254}$$

Performing the substitutions (S), (S') and applying the operator obtained to ψ (denoting, as usual, the potential energy ϵV by U), we get the following equation, which should represent the relativistic generalization of the Schrödinger equation:

$$\left[\frac{1}{c^2}\left(U + \frac{h}{2\pi i}\frac{\partial}{\partial t}\right)^2 - \sum_k\left(\frac{h}{2\pi i}\frac{\partial}{\partial x_k} - \frac{\epsilon}{c}A_k\right)^2 - m_0^2c^2\right]\psi = 0. \tag{255}$$

---

[7] We denote by ϵ the charge of a general particle, since in this chapter we shall reserve the letter e for the *absolute value* of the electronic charge. In the case of the electron, $\epsilon = -e$.

[8] In fact, the kinetic energy is $m_0c^2\left(\frac{1}{\sqrt{1 - v^2/c^2}} - 1\right)$, the rest energy, m_0c^2; the electrostatic energy, ϵV. The total energy is therefore $\frac{m_0c^2}{\sqrt{1 - v^2/c^2}} + \epsilon V$. In order to express the total energy in terms of the p_k, note that we get from (252): $\frac{m_0^2v^2}{1 - v^2/c^2} = \sum_k\left(p_k - \frac{\epsilon}{c}A_k\right)^2$, from which $\frac{m_0^2c^2}{1 - v^2/c^2} = m_0^2c^2 + \sum_k\left(p_k - \frac{\epsilon}{c}A_k\right)^2$. Substituting into the expression for the energy, we get (251).

First we shall verify that in the nonrelativistic limit, that is, when c may be considered sufficiently large compared with the other velocities involved, this equation leads back to the Schrödinger theory. However, we find that the ψ which occurs in (255) differs from the Schrödinger ψ (which we denote by $\bar{\psi}$ for the moment) by a factor of modulus 1, of no importance in the calculation of ρ and **i**, which corresponds to the fact that in the energy there will also be included the rest energy m_0c^2 of the electron. In fact, if we put

$$\psi = \bar{\psi}\, e^{-\frac{2\pi i}{h} m_0 c^2 t}$$

and substitute this expression into (255), we have, by a simple calculation,

$$\left[\frac{1}{c^2}\left(U + \frac{h}{2\pi i}\frac{\partial}{\partial t}\right)^2 - 2m_0\left(U + \frac{h}{2\pi i}\frac{\partial}{\partial t}\right) - \sum_k \left(\frac{h}{2\pi i}\frac{\partial}{\partial x_k} - \frac{\epsilon}{c}A_k\right)^2\right]\bar{\psi} = 0.$$

If we consider the first term to be negligible because of the factor $1/c^2$, we obtain the nonrelativistic approximation

$$\frac{1}{2m_0}\sum_k \left(\frac{h}{2\pi i}\frac{\partial}{\partial x_k} - \frac{\epsilon}{c}A_k\right)^2 \bar{\psi} + U\bar{\psi} = -\frac{h}{2\pi i}\frac{\partial \bar{\psi}}{\partial t},$$

an equation which coincides with (142′) of §31, except for the difference in notation.

However, it was recognized that for various reasons (255) could not in general be adopted as relativistic generalization of the Schrödinger equation; for example, according to (255) the integral of $|\psi|^2$ is varying in time, so that it may not be equated to 1. However, these difficulties may be avoided in the special case of an electron not subject to forces ($V = \mathbf{A} = 0$), in which case the equation becomes

$$\Delta\psi - \frac{1}{c^2}\frac{\partial^2\psi}{\partial t^2} - \frac{4\pi^2 m_0^2 c^2}{h^2}\psi = 0 \tag{256}$$

and constitutes the starting point of the Dirac theory.

47. Principles of the Dirac theory. The fundamental idea which has led to the Dirac theory is as follows. Let us postulate, in analogy to the Schrödinger theory and the Pauli theory, that the probability density P is calculated (as was pointed out in §25 of Part II) by means of a certain, at present undetermined number N of functions $\psi_1, \psi_2, \ldots \psi_N$ (in the Schrödinger theory $N = 1$, in the Pauli theory $N = 2$; we shall see in what follows that in the Dirac theory we have to take $N = 4$), that is, by the relation

$$P = \psi_1\psi_1^* + \psi_2\psi_2^* + \psi_3\psi_3^* + \cdots + \psi_N\psi_N^*. \tag{257}$$

If each of the ψ were contained only in a differential equation of the second order in the time, we should have to assign the initial values of ψ and $\partial\psi/\partial t$ in order to determine the further evolution of ψ in time. Instead, it is obvious to postulate that the assignment of the initial value of ψ (in all space) is sufficient to determine P at all later instants, and hence to assume that the N functions ψ *satisfy a system of N differential equations of the first order in the time.* Since in any relativistic theory the variable t must be treated in the same manner as the spatial coordinates x_1, x_2, x_3, it follows that these equations must also be of the first order with respect to x_1, x_2, x_3. Of course, by differentiation it is always possible to obtain second-order equations from equations of the first order—equations which will be necessary consequences of the first (but not conversely). We shall therefore require, in the particular case of no electromagnetic field, that the relativistic equation (256) be satisfied by each of the ψ, as a consequence of the first-order equations which we are about to establish.[9] At first we shall confine ourselves to the case of a free electron.

The simplest hypothesis which might be made concerning the desired equations is that they be linear with constant coefficients.[10] As we shall see, it is possible to satisfy all required conditions by this assumption. On the supposition that the equations may be solved for the time derivatives, the system may be written in the form

$$\frac{1}{c}\frac{\partial\psi_\mu}{\partial t} + \sum_{\lambda,k} \alpha_{\mu\lambda}{}^k \frac{\partial\psi_\lambda}{\partial x_k} + \frac{2\pi i}{h}\sum_\lambda \beta_{\mu\lambda}\psi_\lambda = 0 \tag{258}$$

---

[9] Analogously, in the electromagnetic theory of light, the wave equation (of the second order) is a consequence of the Maxwell equations (of the first order).

[10] This hypothesis may also be justified by the argument that no point of space-time should occupy a privileged position.

(where $\mu = 1, 2, \ldots N$), in which the coefficients $\alpha_{\mu\lambda}{}^k$, $\beta_{\mu\lambda}$ are constants to be determined. (The reason for the factor $2\pi i/h$ evident in the last summation will appear subsequently.) Furthermore, we have introduced the convention, which will be retained in the following, of denoting by Greek letters $\lambda, \mu, \ldots$ the indices (taking the values from 1 to N) which distinguish the various ψ, and by ordinary letters the indices ($= 1, 2, 3$) which distinguish the three space coordinates (which are designated indifferently by x, y, z or x_1, x_2, x_3). The writing is considerably simplified if we introduce matrix notation, denoting by α^k and β, respectively, the four matrices with N rows and N columns, whose element of row μ and column λ is $\alpha_{\mu\lambda}{}^k$, $\beta_{\mu\lambda}$, and consider ψ as the symbol of a matrix with N rows and a single column, as was done in §45 for $N = 2$. Then the N equations (258) may be condensed into the expression

$$\frac{1}{c}\frac{\partial\psi}{\partial t} + \sum_k \alpha^k \frac{\partial\psi}{\partial x_k} + \frac{2\pi i}{h}\beta\psi = 0, \tag{259}$$

which, upon the introduction of the operators

$$\mathfrak{p}_k = \frac{h}{2\pi i}\frac{\partial}{\partial x_k}, \qquad \mathfrak{p}_4 = -\frac{1}{c}\frac{h}{2\pi i}\frac{\partial}{\partial t}$$

is more conveniently written

$$\left[-\mathfrak{p}_4 + \beta + \sum_k \alpha^k\mathfrak{p}_k\right]\psi = 0. \tag{259'}$$

Before proceeding to the determination of the coefficients of (258), that is, of the four matrices α^k, β, we should obtain the expressions for the average electric charge density ρ and average electric current density $\mathbf{j}$, which will be the generalizations of those already found for the nonrelativistic case.

48. Electric charge density and electric current density. The probability density P is given by (257). Upon introducing the matrix ψ and also the one-rowed matrix with N columns

$$\tilde{\psi} = |\psi_1^*, \psi_2^*, \ldots, \psi_N^*| \tag{260}$$

(notation in accordance with that of §7), we may write (257) in the short form

$$P = \tilde{\psi}\psi, \tag{261}$$

as may readily be verified by applying the rule for the product of two matrices. Consequently, the average electric charge density $\rho = -eP$ will be given by

$$\rho = -e\tilde{\psi}\psi. \tag{262}$$

In order to find the average electric current density **j**, we observe that the latter must satisfy the "equation of continuity"

$$\frac{\partial \rho}{\partial t} + \operatorname{div} \mathbf{j} = 0, \tag{263}$$

which, when the expression for ρ is introduced, becomes

$$\operatorname{div} \mathbf{j} = e\left(\frac{\partial \tilde{\psi}}{\partial t}\psi + \tilde{\psi}\frac{\partial \psi}{\partial t}\right).$$

In the last equation, let us replace the derivatives of ψ and $\tilde{\psi}$ by their expressions obtained from (259) and from its complex conjugate. Denoting, as usual, by $\tilde{\alpha}^k$, $\tilde{\beta}$ the matrices obtained from α^k and β by interchanging rows and columns, and taking the complex conjugate of each element,[11] we obtain

$$\frac{1}{c}\frac{\partial \tilde{\psi}}{\partial t} + \sum_k \frac{\partial \tilde{\psi}}{\partial x_k}\tilde{\alpha}_k - \frac{2\pi i}{h}\tilde{\psi}\tilde{\beta} = 0.$$

Substituting these derivatives into the expression for div **j**, we get

$$\operatorname{div} \mathbf{j} = -ce\sum_k \left(\frac{\partial \tilde{\psi}}{\partial x_k}\tilde{\alpha}^k\psi + \tilde{\psi}\alpha^k\frac{\partial \psi}{\partial x_k}\right) - \frac{2\pi i}{h}ce\tilde{\psi}(\beta - \tilde{\beta})\psi.$$

In order that the right-hand member actually have the form of a divergence, it suffices to impose upon the matrices the conditions

$$\tilde{\alpha}^k = \alpha^k, \qquad \tilde{\beta} = \beta,$$

which imply that these matrices must be Hermitian. The formula then becomes

$$\operatorname{div} \mathbf{j} = -ce\sum_k \frac{\partial}{\partial x_k}(\tilde{\psi}\alpha^k\psi),$$

---

[11] Note that in order to preserve the validity of the rule for matrix multiplication, the matrix ψ is always written to the right of α^k, β, and $\tilde{\psi}$ to the left.

and we may take as components of **j** the expressions

$$j_k = -ce\tilde{\psi}\alpha^k\psi, \tag{264}$$

or else explicitly,

$$j_k = -ce\sum_{\lambda,\mu}\psi_\lambda{}^*\alpha_{\lambda\mu}{}^k\psi_\mu. \tag{264'}$$

49. The Dirac equations for the free electron. Let us now determine the coefficients of equations (258), or the matrices α^k, β, imposing the condition that from these first-order equations there follows, as a necessary consequence, the relativistic equation (256) for each of the N functions [or that equation (256) be satisfied also when ψ is considered as a matrix]. For this purpose, let us apply to (259) the following differential operator, which is the only one for which the terms in $\partial/\partial t$ disappear:

$$-\frac{1}{c}\frac{\partial}{\partial t} + \sum_k \alpha^k\frac{\partial}{\partial x_k} + \frac{2\pi i}{h}\beta.$$

Taking into account the fact that the matrices α^k, β commute with the differentiation symbols, but are not in general to be considered as commuting with each other, we obtain

$$-\frac{1}{c^2}\frac{\partial^2\psi}{\partial t^2} + \sum_{j,k}\alpha^k\alpha^j\frac{\partial^2\psi}{\partial x_k\partial x_j} + \frac{2\pi i}{h}\sum_k(\beta\alpha^k + \alpha^k\beta)\frac{\partial\psi}{\partial x_k} - \frac{4\pi^2}{h^2}\beta^2\psi = 0.$$

In order that this equation be identical with (256), we must have

$$\begin{aligned} \alpha^k\alpha^j + \alpha^j\alpha^k &= 0, \qquad \text{(for } j \neq k) \\ (\alpha^k)^2 &= \{1\}, \\ \beta\alpha^k + \alpha^k\beta &= 0, \\ \beta^2 &= m_0^2c^2\{1\}, \end{aligned}$$

where $\{1\}$ stands for the unit matrix with N rows and N columns. Introducing, instead of β, the matrix α^4 defined by

$$\beta = m_0c\alpha^4, \tag{265}$$

we may condense the preceding formulas as follows:

$$\left.\begin{aligned} \alpha^\lambda\alpha^\mu + \alpha^\mu\alpha^\lambda &= 0, \\ (\alpha^\lambda)^2 &= \{1\}, \end{aligned}\right\} \qquad \text{(for } \lambda \neq \mu) \tag{266}$$

where the indices λ and μ, according to the convention already adopted, take the values 1, 2, . . . N. We may also write, gathering (266) into a single formula,

$$\alpha^\lambda \alpha^\mu + \alpha^\mu \alpha^\lambda = 2\delta_{\lambda\mu}\{1\}. \tag{266'}$$

The problem is therefore to find four Hermitian matrices which satisfy these conditions. It may be shown that it is not possible to find such matrices with rank less than 4. However, it is possible to find them (in an infinite number of ways) with $N \geq 4$.[12] Since we want to limit the complication of the theory as much as possible, we shall assume $N = 4$; that is, we shall represent the state of an electron by four functions ψ_1, ψ_2, ψ_3, ψ_4 of x, y, z, t, or else by the matrix

$$\psi = \begin{vmatrix} \psi_1 \\ \psi_2 \\ \psi_3 \\ \psi_4 \end{vmatrix}.$$

As far as the four matrices α are concerned, we may take the following, which, as may be verified, are Hermitian and satisfy (266′):

$$\left.\begin{array}{ll}
\alpha^1 = \begin{vmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{vmatrix} & \alpha^2 = \begin{vmatrix} 0 & 0 & 0 & -i \\ 0 & 0 & i & 0 \\ 0 & -i & 0 & 0 \\ i & 0 & 0 & 0 \end{vmatrix} \\[2ex]
\alpha^3 = \begin{vmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{vmatrix} & \alpha^4 = \begin{vmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{vmatrix}.
\end{array}\right\} \tag{267}$$

The fundamental equation (259′), which, when (265) is introduced for β, becomes

$$\left[\sum_k \alpha^k \mathfrak{p}_k - \mathfrak{p}_4 + m_0 c \alpha^4\right] \psi = 0, \tag{268}$$

---

[12] And precisely for all values of N that are multiples of 4. Such solutions, however, may be reduced to the solution for $N = 4$.

is equivalent, when we take expressions (267) for the α, to the following four equations (Dirac equations for the free electron):

$$\left.\begin{aligned}(-\mathfrak{p}_4 + m_0c)\psi_1 + (\mathfrak{p}_1 - i\mathfrak{p}_2)\psi_4 + \mathfrak{p}_3\psi_3 &= 0,\\ (-\mathfrak{p}_4 + m_0c)\psi_2 + (\mathfrak{p}_1 + i\mathfrak{p}_2)\psi_3 - \mathfrak{p}_3\psi_4 &= 0,\\ (-\mathfrak{p}_4 - m_0c)\psi_3 + (\mathfrak{p}_1 - i\mathfrak{p}_2)\psi_2 + \mathfrak{p}_3\psi_1 &= 0,\\ (-\mathfrak{p}_4 - m_0c)\psi_4 + (\mathfrak{p}_1 + i\mathfrak{p}_2)\psi_1 - \mathfrak{p}_3\psi_2 &= 0.\end{aligned}\right\} \qquad (269)$$

These may also be written, with the explicit expressions for the operators,

$$\left(\frac{1}{c}\frac{\partial}{\partial t} + \frac{2\pi i m_0 c}{h}\right)\psi_1 + \left(\frac{\partial}{\partial x_1} - i\frac{\partial}{\partial x_2}\right)\psi_4 + \frac{\partial \psi_3}{\partial x_3} = 0,$$

and so on.

The conditions (266′) may also be satisfied by an infinity of other groups of four Hermitian matrices; there are thus as many different forms of the Dirac equations as there are different groups of four functions ψ. First of all, we see immediately that (266′) is also satisfied by the matrices α'^λ which are obtained from the α^λ defined by relations (267), by the transformation $\alpha'^\lambda = S\alpha^\lambda S^{-1}$, where S is an arbitrary transformation matrix (provided that it is "unitary"). Then (268) becomes

$$\left[-\mathfrak{p}_4 + m_0cS\alpha^4S^{-1} + \sum_k S\alpha^kS^{-1}\mathfrak{p}_k\right]\psi' = 0,$$

where the new quadruplet of functions ψ is indicated by ψ'. Evidently this equation is satisfied upon taking $\psi' = S\psi$. Note that if we apply (262) and (264) and start from ψ', we obtain the same values for ρ and $\mathbf{j}$ as when starting from ψ. Hence the solution under consideration is not physically different from the previous one. It may be shown that no other quadruplets of matrices (with $N = 4$) satisfy the desired conditions.

50. The electron in an electromagnetic field. Now we have to extend the Dirac equations to the case of an electron situated in an electric and magnetic field, derivable from a scalar (electrostatic) potential V and from a vector potential $\mathbf{A}$. The extension is analogous to that of nonrelativistic wave mechanics, in which the equation for a particle in a magnetic field [see formula (142′), §31] differs from that in the absence of a field in having, instead of the operators $\frac{h}{2\pi i}\frac{\partial}{\partial q_r}$, $-\frac{h}{2\pi i}\frac{\partial}{\partial t}$, the operators $\frac{h}{2\pi i}\frac{\partial}{\partial q_r} - \frac{e}{c}A_r$, $-\frac{h}{2\pi i}\frac{\partial}{\partial t} - U$. Recalling that now the charge of the electron is indicated by $-e$ and that $U = -eV$, we are led to replace, in (268), the operators $\mathfrak{p}_k$ and $\mathfrak{p}_4$, respectively, by

$$\mathfrak{P}_k = \frac{h}{2\pi i}\frac{\partial}{\partial x_k} + \frac{e}{c}A_k, \qquad \mathfrak{P}_4 = -\frac{h}{2\pi i}\frac{1}{c}\frac{\partial}{\partial t} + \frac{e}{c}V. \tag{270}$$

This substitution is suggested not only by the said analogy, but also by the necessity of preserving the relativistic invariance of the equation (in fact, in a Lorentz transformation A_k and V transform in the same way as the operators $\partial/\partial x_k$, $\partial/\partial t$) as well as its "gauge invariance," which will be explained in §57. The Dirac equation then becomes

$$\left[\sum_k \alpha^k\mathfrak{P}_k - \mathfrak{P}_4 + m_0c\alpha^4\right]\psi = 0, \tag{271}$$

which, assuming expressions (267) for the α, may be translated into the four equations

$$\left.\begin{aligned}(-\mathfrak{P}_4 + m_0c)\psi_1 + (\mathfrak{P}_1 - i\mathfrak{P}_2) + \mathfrak{P}_3\psi_3 &= 0,\\ (-\mathfrak{P}_4 + m_0c)\psi_2 + (\mathfrak{P}_1 + i\mathfrak{P}_2) - \mathfrak{P}_3\psi_4 &= 0,\\ (-\mathfrak{P}_4 - m_0c)\psi_3 + (\mathfrak{P}_1 - i\mathfrak{P}_2) + \mathfrak{P}_3\psi_1 &= 0,\\ (-\mathfrak{P}_4 - m_0c)\psi_4 + (\mathfrak{P}_1 + i\mathfrak{P}_2) - \mathfrak{P}_3\psi_2 &= 0.\end{aligned}\right\} \tag{272}$$

Of course, the justification of this postulate lies in the consequences which are deduced from it: in particular, in the fact that the results of these equations coincide, in first approximation, with the results of the Pauli theory and, if the spin is neglected, with the results of the Schrödinger theory; they also strictly satisfy the relativity principle (as distinct from the other mentioned theories). All this will be seen in the following sections. It is now to be noted that if equation (271) is solved for the time derivative contained in $\mathfrak{P}_4$, it may be written in the usual form:

$$\mathfrak{H}\psi = -\frac{h}{2\pi i}\frac{\partial\psi}{\partial t}, \tag{273}$$

if we put

$$\mathfrak{H} = c\sum_k \alpha^k \mathfrak{P}_k - eV + m_0c^2\alpha^4. \tag{274}$$

This operator may thus be considered to be the Hamiltonian operator of the Dirac theory. It is apparent that from (273) we should obtain formula (118) for the derivative of an observable by the same procedure as used in §28. This relation may therefore be taken to hold also in the Dirac theory, provided that we understand by $\mathfrak{H}$ the operator (274).

Finally, we note that for a stationary state of energy[13] W we have

$$\psi_\lambda = u_\lambda\, e^{-\frac{2\pi i}{h}Wt},$$

and the matrix composed of the four u satisfies (just as does ψ in this case) the equation obtained from (271) when replacing $\mathfrak{P}_4$ by $(1/c)(W + eV)$.

51. The Pauli theory as a first approximation. We are now going to show how the Dirac equations reduce to the equations of the Pauli theory when terms in $1/c^2$ are neglected. For this purpose, it is convenient to introduce, instead of the four ψ, two pairs of functions φ_1, φ_2, and χ_1, χ_2, related to the ψ by the following expressions:

$$\left.\begin{aligned} \psi_1 &= \varphi_1\, e^{-\frac{2\pi i}{h}m_0c^2t}, & \psi_2 &= \varphi_2\, e^{-\frac{2\pi i}{h}m_0c^2t}, \\ \psi_3 &= \chi_1\, e^{-\frac{2\pi i}{h}m_0c^2t}, & \psi_4 &= \chi_2\, e^{-\frac{2\pi i}{h}m_0c^2t}. \end{aligned}\right\} \tag{275}$$

Applying the operators $\mathfrak{P}$ defined in §50, we immediately find the formulas

$$\mathfrak{P}_k\left|\begin{matrix}\psi_1\\ \psi_2\\ \psi_3\\ \psi_4\end{matrix}\right| = e^{-\frac{2\pi i}{h}m_0c^2t}\,\mathfrak{P}_k\left|\begin{matrix}\varphi_1\\ \varphi_2\\ \chi_1\\ \chi_2\end{matrix}\right|, \qquad \mathfrak{P}_4\left|\begin{matrix}\psi_1\\ \psi_2\\ \psi_3\\ \psi_4\end{matrix}\right| = e^{-\frac{2\pi i}{h}m_0c^2t}(\mathfrak{P}_4 + m_0c)\left|\begin{matrix}\varphi_1\\ \varphi_2\\ \chi_1\\ \chi_2\end{matrix}\right|.$$

Taking the last of these into account, we see that in the first two equations of (272) the term m_0c disappears from the first

---

[13] By W we indicate the total energy; that is, we include the rest energy, required by relativity theory, equal to m_0c^2. Hence between W and the E used thus far the relation $W = E + m_0c^2$ holds.

parenthesis, while the same term is doubled in the other two. In fact, the equations become

$$\left.\begin{aligned} -\mathfrak{P}_4\varphi_1 + (\mathfrak{P}_1 - i\mathfrak{P}_2)\chi_2 + \mathfrak{P}_3\chi_1 &= 0,\\ -\mathfrak{P}_4\varphi_2 + (\mathfrak{P}_1 + i\mathfrak{P}_2)\chi_1 - \mathfrak{P}_3\chi_2 &= 0,\\ (-\mathfrak{P}_4 - 2m_0c)\chi_1 + (\mathfrak{P}_1 - i\mathfrak{P}_2)\varphi_2 + \mathfrak{P}_3\varphi_1 &= 0,\\ (-\mathfrak{P}_4 - 2m_0c)\chi_2 + (\mathfrak{P}_1 + i\mathfrak{P}_2)\varphi_1 - \mathfrak{P}_3\varphi_2 &= 0. \end{aligned}\right\} \tag{276}$$

Let us introduce the two-rowed matrices $\varphi = \left|\begin{matrix}\varphi_1\\ \varphi_2\end{matrix}\right|$, $\chi = \left|\begin{matrix}\chi_1\\ \chi_2\end{matrix}\right|$, and in addition, the three matrices (with two rows and two columns) σ_x, σ_y, σ_z, defined in §45, which for convenience we shall now designate by σ_1, σ_2, σ_3. Then the four preceding equations may be assembled into the formulas

$$\mathfrak{P}_4\varphi = \sum_k \mathfrak{P}_k\sigma_k\chi, \tag{277}$$

$$(\mathfrak{P}_4 + 2m_0c)\chi = \sum_k \mathfrak{P}_k\sigma_k\varphi. \tag{278}$$

In ordinary cases (that is, those corresponding, in the classical model, to particles having velocities small compared with c, so that nonrelativistic mechanics may be used), $\mathfrak{P}_4\chi$ turns out to be negligible compared with $2m_0c\chi$, so that we obtain from (278) the following approximate expression for χ:

$$\chi = \frac{1}{2m_0c}\sum_k \mathfrak{P}_k\sigma_k\varphi, \tag{278$'$}$$

from which it may be seen that the χ are ordinarily small compared with the φ (or ψ_3 and ψ_4 are small compared with ψ_1 and ψ_2). Substituting this expression of χ into (277), we have

$$\mathfrak{P}_4\varphi = \frac{1}{2m_0c}\sum_{k,l} \mathfrak{P}_k\mathfrak{P}_l\sigma_k\sigma_l\varphi. \tag{277$'$}$$

In this double summation, the six terms for which $k \neq l$ may be united two by two in the following manner. For instance, let us consider the two terms $(\mathfrak{P}_2\mathfrak{P}_3\sigma_2\sigma_3 + \mathfrak{P}_3\mathfrak{P}_2\sigma_3\sigma_2)$. By virtue of (236), they may be written $(\mathfrak{P}_2\mathfrak{P}_3 - \mathfrak{P}_3\mathfrak{P}_2)i\sigma_1\varphi$. On the other hand, we readily perceive from (270) that

$$(\mathfrak{P}_2\mathfrak{P}_3 - \mathfrak{P}_3\mathfrak{P}_2)\varphi = \frac{h}{2\pi i}\frac{e}{c}\left(\frac{\partial A_3}{\partial x_2} - \frac{\partial A_2}{\partial x_3}\right)\varphi = \frac{h}{2\pi i}\frac{e}{c}H_1\varphi,$$

so that the two terms in question reduce to $\frac{h}{2\pi}\frac{e}{c}H_1\sigma_1\varphi$, and the entire group of six terms with $k \neq l$ gives

$$\frac{eh}{2\pi c}\boldsymbol{\sigma}\cdot\mathbf{H}\varphi.$$

As for the three terms in which $k = l$, they give, if we keep (234) in mind,

$$\sum_k \mathfrak{P}_k^2\varphi.$$

Thus finally (277′) becomes

$$\mathfrak{P}_4\varphi = \left[\frac{eh}{4\pi m_0c^2}\boldsymbol{\sigma}\cdot\mathbf{H} + \frac{1}{2m_0c}\mathfrak{P}_k^2\right]\varphi,$$

or also, when we write $\mathfrak{P}_4$ explicitly and recall the expression of the Bohr magneton $\mu_0 = \frac{eh}{4\pi mc}$,

$$-\frac{h}{2\pi i}\frac{\partial\varphi}{\partial t} = \left[\frac{1}{2m_0}\sum_k \mathfrak{P}_k^2 - eV + \mu_0\boldsymbol{\sigma}\cdot\mathbf{H}\right]\varphi. \tag{279}$$

The square-bracket expression on the right side of this equation may be identified with the operator $\mathfrak{H}$ of (244), and hence this equation is identical with that in the Pauli theory.

From this we see to what extent the Dirac equations imply, at least in first approximation, the existence of a magnetic moment equal to μ_0 directed in the sense opposite to the spin. This consequence will be seen in a different manner in the following section.

52. Magnetic moment of the electron. Let us now show that the Dirac electron behaves (in first approximation) as if it had a magnetic moment μ_0, not only in regard to the energy levels, as seen in the last section, but also inasmuch as it generates an average magnetic field[14] equal to the field produced by a magnetic dipole

---

[14] In the sense explained in §27.

of moment μ_0. For simplicity we shall confine ourselves here to the case of an electron at rest, possibly located in an electric field (but not in a magnetic field of external origin),[15] having its spin parallel to the z-axis, so that in the Pauli approximation $\psi_2 = \varphi_2 = 0$.

First let us calculate the components of the average current density j (from which we must then obtain the average magnetic field generated by the electron) by means of equation (264), introducing the φ and χ defined in the previous section by (275). Using (267), we get

$$\begin{aligned} j_1 &= -ce(\varphi_1^*\chi_2 + \chi_2^*\varphi_1), \\ j_2 &= -ice(-\varphi_1^*\chi_2 + \chi_2^*\varphi_1), \\ j_3 &= -ce(\varphi_1^*\chi_1 + \chi_1^*\varphi_1). \end{aligned}$$

Let us now introduce the approximate expression (278′) for the χ. In the actual case, if we make use of (241) and (241′), this becomes

$$\chi = \frac{h}{4\pi i m_0 c} \left| \begin{matrix} \dfrac{\partial \varphi_1}{\partial z} \\ \dfrac{\partial \varphi_1}{\partial x} + i\dfrac{\partial \varphi_1}{\partial y} \end{matrix} \right|.$$

In addition, we recall that φ_1 satisfies the ordinary Schrödinger equation (see page 154) and that, for a stationary state of energy E, the Schrödinger equation always has a solution of the form $\varphi_1 = u\, e^{-\frac{2\pi i}{h}Et}$, with u real (see page 162). We then obtain

$$j_1 = -c\mu_0\frac{\partial u^2}{\partial y}, \qquad j_2 = c\mu_0\frac{\partial u^2}{\partial x}, \qquad j_3 = 0.$$

Upon introduction of the vector **I** with components

$$I_1 = 0, \qquad I_2 = 0, \qquad I_3 = -\mu_0 u^2, \tag{280}$$

these formulas may be assembled into the vectorial expression

$$\mathbf{j} = c \operatorname{curl} \mathbf{I}. \tag{281}$$

---

[15] For a more general treatment, see, for instance, No. 26 of the Bibliography, Chapters XIII and XIV. With the more rigorous theory we find [see G. Breit, *Nature* **122,** 649 (1928)] that an electron bound to a nucleus of charge Ze has, in the ground state, a magnetic moment $\frac{1}{3}(1 + 2\sqrt{1 - \alpha^2 Z^2})\,\mu_0$, where α is the fine-structure constant.

Hence the current density does not turn out to be zero, as it would have been in the Schrödinger theory for a ψ of the form considered here. However, the total current across any infinite plane vanishes as may immediately be verified. This statement may be interpreted in terms of a model by saying that the electric charge density "cloud," which is equivalent to the electron on the average, does not displace itself as a whole but contains internal electric currents, whose magnetic effects we shall now investigate.

The vector potential, from which the magnetic field is derived, is obtained from the current density $\mathbf{j}$ by the well-known formula from electromagnetic theory

$$\mathbf{A} = \frac{1}{c}\int_S \frac{\mathbf{j}}{r}\,dS. \tag{282}$$

On the other hand, it may be shown[16] that a magnetized body whose intensity of magnetization is $\mathbf{I}$, gives rise to a magnetic field whose vector potential is given by

$$\mathbf{A} = \int_S \frac{\operatorname{curl}\mathbf{I}}{r}\,dS.$$

By comparing these two expressions and taking (281) into account, we may conclude that the magnetic effect of the currents in question is the same as would be produced by an intensity of magnetization of space represented by the vector $\mathbf{I}$, defined by (280). In other words, the "cloud" of electric charge density behaves as if it were magnetized, in the direction of the *negative* z-axis (that is, in the sense opposite to the angular momentum) with an intensity $\mu_0 u^2$. The total magnetic moment which is obtained by integrating this intensity over all space, proves to be μ_0, by virtue of the normalization of u.

If we had supposed instead that the spin was antiparallel to the z-axis, that is, $\varphi_1 = 0$ and $\varphi_2 = u\,e^{-\frac{2\pi i}{h}Et}$, we should have obtained an analogous result, but the magnetic moment would have been directed along the positive z-axis, that is, opposite to the spin in that case also.

---

[16] See, for instance, No. 26 of the Bibliography, page 173.

In the general case we find that the equivalent magnetization is given to the same approximation by

$$\left\{\begin{aligned} I_1 &= -\mu_0(\psi_1^*\psi_2 + \psi_2^*\psi_1), \\ I_2 &= i\mu_0(\psi_1^*\psi_2 - \psi_2^*\psi_1), \\ I_3 &= -\mu_0(\psi_1^*\psi_1 - \psi_2^*\psi_2). \end{aligned}\right\} \tag{283}$$

We might suppose that we could detect the magnetic moment of the electron experimentally (as is done for atoms) by means of an experiment of the Stern-Gerlach type. This is not possible because, no matter what the experimental arrangement, the spreading of the electron beam due to diffraction is of the same order of magnitude as the deviation due to the magnetic forces we are trying to detect.[17] In fact, Bohr has pointed out that to the order of approximation in which wave mechanics coincides with point mechanics, the spin and the magnetic moment of the electron (which both occur in terms containing h) disappear; hence no experiment in which the corpuscular model may be applied to the electron permits the detection of these moments.

However, it is not impossible to imagine experiments of the non-corpuscular type which would allow us to prove the existence of the spin. For example, a beam of electrons diffracted by a crystal in a direction nearly perpendicular to the incident beam should presumably be "polarized"; that is to say, if it is subsequently diffracted by another crystal, it should exhibit different degrees of diffraction in the various planes passing through its direction of propagation. (In the vector model, this change might be interpreted by imagining that after the first diffraction the spins of the electrons are no longer oriented at random but with a preference for one of these planes.) This experiment, performed in different ways by various investigators, has given results which are still open to question.

53. The spin in the Dirac equations. In order to show that (271) implicitly contains the existence of spin and its property of quantization (which also results from §51, but only to a first approximation), let us consider the case in which the force acting upon the electron has zero moment with respect to the z-axis, such as for a central field. In this case we have shown (see §30) that in the Schrödinger theory the angular momentum about the z-axis remains constant, or that

$$M_z = xp_y - yp_x \tag{284}$$

is a first integral, as in ordinary mechanics. We shall now show that in the Dirac theory this is no longer so, and that instead of M_z, the observable $N_z = M_z + S_z$ will be a constant, where S_z is an

---

[17] For the proof, see N. F. Mott, *Proc. Roy. Soc.* **A124**, 425 (1929).

observable whose eigenvalues are $\pm h/4\pi$. This result signifies that the total angular momentum with respect to the z-axis is not M_z but N_z, and that the term S_z represents an intrinsic angular momentum of the electron, whose projection upon the z-axis is always $\pm h/4\pi$.

In order to realize the above-mentioned conditions, let us suppose that $\mathbf{A} = 0$ (no magnetic field) and that V is symmetric about the z-axis; that is if we adopt polar coordinates and if φ is the "longitude," V is to be independent of φ. Let us now calculate the derivative of M_z with respect to t, using (118) of §28. We get (ignoring the distinction between the symbol for an observable and the symbol of its operator)

$$\dot{M}_z = \frac{2\pi i}{h}(\mathcal{H}M_z - M_z\mathcal{H}). \tag{285}$$

It is now apparent that in the present case the Hamiltonian reduces to [see equation (274)]

$$\mathcal{H} = c\sum_k \alpha^k p_k - eV + m_0c^2\alpha^4, \tag{286}$$

and that M_z, by (284), commutes with α^1, α^2, α^3, α^4, with p_z, and also with V (since, as we have seen in §30, $M_z = (h/2\pi i)(\partial/\partial\varphi)$ in polar coordinates, and V is independent of φ). Hence in (285) there remains only the contribution of the first two terms in the summation of (286):

$$\dot{M}_z = \frac{2\pi i}{h}[c\alpha^1(p_xM_z - M_zp_x) + c\alpha^2(p_yM_z - M_zp_y)].$$

But from the commutation relations and from (284) we obtain

$$p_xM_z - M_zp_x = \frac{h}{2\pi i}p_y, \qquad p_yM_z - M_zp_y = -\frac{h}{2\pi i}p_x;$$

hence

$$\dot{M}_z = c(\alpha^1p_y - \alpha^2p_x). \tag{287}$$

As we see, this derivative is not identically zero, which means that M_z is not a first integral.

Let us now consider the observable S_z whose operator is

$$S_z = \frac{h}{4\pi i}\alpha^1\alpha^2, \tag{288}$$

and let us calculate its derivative by the formula used above. We have

$$\dot S_z = \tfrac{1}{2}(\mathcal{H}\alpha^1\alpha^2 - \alpha^1\alpha^2\mathcal{H}).$$

Substituting for $\mathcal{H}$ the expression (286), we note that $\alpha^1\alpha^2$ commutes with the p and with V, and that furthermore we have, as an immediate consequence of (266),

$$\begin{aligned}
\alpha^1(\alpha^1\alpha^2) - (\alpha^1\alpha^2)\alpha^1 &= 2\alpha^2,\\
\alpha^2(\alpha^1\alpha^2) - (\alpha^1\alpha^2)\alpha^2 &= -2\alpha^1,\\
\alpha^3(\alpha^1\alpha^2) - (\alpha^1\alpha^2)\alpha^3 &= 0,\\
\alpha^4(\alpha^1\alpha^2) - (\alpha^1\alpha^2)\alpha^4 &= 0.
\end{aligned}$$

Thus we find

$$\dot S_z = c(\alpha^2 p_x - \alpha^1 p_y).$$

Comparing this expression with (287), we find that

$$\dot M_z + \dot S_z = 0,]$$

that is, that the observable $N_z = M_z + S_z$ is a first integral, as was claimed. Similar reasoning may be applied for the y and z components. We conclude that the projections of the spin upon the three axes are represented by the operators

$$S_x = \frac{h}{2\pi i}\,\alpha^2\alpha^3, \qquad S_y = \frac{h}{2\pi i}\,\alpha^3\alpha^1, \qquad S_z = \frac{h}{2\pi i}\,\alpha^1\alpha^2,$$

or, adopting the matrices (267) and carrying out the products, by

$$S_x = \frac{h}{4\pi}\begin{vmatrix} 0 & 1 & 0 & 0\\ 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 1\\ 0 & 0 & 1 & 0\end{vmatrix}, \qquad S_y = \frac{h}{4\pi}\begin{vmatrix} 0 & -i & 0 & 0\\ i & 0 & 0 & 0\\ 0 & 0 & 0 & -i\\ 0 & 0 & i & 0\end{vmatrix},$$

$$S_z = \frac{h}{4\pi}\begin{vmatrix} 1 & 0 & 0 & 0\\ 0 & -1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & -1\end{vmatrix}. \tag{289}$$

It is to be noted first of all that these matrices are Hermitian, and hence the observables which they represent are real.[18] Further-

---

[18] The particular choice adopted for the matrices makes S_z diagonal. This means that the operators are in the "S_z-representation." By a transformation of the type described on page 408, we could pass to a representation in which S_y or S_x would be diagonal.

more, each of these has the eigenvalues $\pm h/4\pi$. In fact, it may readily be verified that the three matrices written above have their squares equal to the unit matrix, and hence that S_x^2, S_y^2, S_z^2 have the single eigenvalue $(h/4\pi)^2$. It is also apparent that S_x, S_y, S_z do not commute with one another.[19] Instead, they satisfy the same commutation relations which were found in §30 for ordinary angular momenta.

Of course, all the considerations of Chapter 11 for operators corresponding to other observables apply to the spin operators. The only difference is that the latter operate upon functions of a variable with only four values, namely, the index λ of the four $\psi_\lambda(x, y, z)$. (Since they do not involve the other variables x, y, z, we are dealing with "incomplete" operators.) These operators are therefore simply linear transformations of groups of four numbers. Correspondingly, each of these has only four eigenvalues and four eigenfunctions (in fact, the four eigenvalues reduce to two double eigenvalues).

In order to illustrate these matters by an example, let us seek the eigenvalues and eigenfunctions of the operator S_z defined by (288) or by the last expression of (289). We indicate by φ^μ ($\mu = 1$, 2, 3, 4) a general one of the four eigenfunctions (which will reduce to a set of four numbers $\varphi_1{}^\mu$, $\varphi_2{}^\mu$, $\varphi_3{}^\mu$, $\varphi_4{}^\mu$), and by s^μ the corresponding eigenvalue. We must then have

$$S_z\varphi^\mu = s^\mu\varphi^\mu,$$

or, in explicit form,

$$\frac{h}{4\pi}\begin{vmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{vmatrix} \cdot \begin{vmatrix} \varphi_1{}^\mu \\ \varphi_2{}^\mu \\ \varphi_3{}^\mu \\ \varphi_4{}^\mu \end{vmatrix} = s^\mu \begin{vmatrix} \varphi_1{}^\mu \\ \varphi_2{}^\mu \\ \varphi_3{}^\mu \\ \varphi_4{}^\mu \end{vmatrix}, \qquad \text{or:} \qquad \frac{h}{4\pi}\begin{vmatrix} \varphi_1{}^\mu \\ -\varphi_2{}^\mu \\ \varphi_3{}^\mu \\ -\varphi_4{}^\mu \end{vmatrix} = s^\mu \begin{vmatrix} \varphi_1{}^\mu \\ \varphi_2{}^\mu \\ \varphi_3{}^\mu \\ \varphi_4{}^\mu \end{vmatrix},$$

which is equivalent to the four equations

$$\left(s^\mu - \frac{h}{4\pi}\right)\varphi_1{}^\mu = 0, \qquad \left(s^\mu + \frac{h}{4\pi}\right)\varphi_2{}^\mu = 0,$$

$$\left(s^\mu - \frac{h}{4\pi}\right)\varphi_3{}^\mu = 0, \qquad \left(s^\mu + \frac{h}{4\pi}\right)\varphi_4{}^\mu = 0.$$

---

[19] Hence the three spin components are not compatible observables. It is because of this fact that the properties of the spin do not correspond at all to those of an ordinary gyroscope.

Since we have to exclude the solution $\varphi_1{}^\mu = \varphi_2{}^\mu = \varphi_3{}^\mu = \varphi_4{}^\mu = 0$, we must have either $s = -h/4\pi$ or $s = +h/4\pi$. Since each of these eigenvalues is double, as we shall see shortly, we shall count them twice and put

$$s^1 = s^3 = +\frac{h}{4\pi}, \qquad s^2 = s^4 = -\frac{h}{4\pi}.$$

In the first case ($\mu = 1, 3$), the equations yield $\varphi_2{}^\mu = \varphi_4{}^\mu = 0$, whereas $\varphi_1{}^\mu$ and $\varphi_3{}^\mu$ remain arbitrary (except for orthonormality) and may be taken equal to 1 and 0, respectively (for $\mu = 1$), or to 0 and 1 (for $\mu = 3$). Similarly, in the second case ($\mu = 2, 4$), we have $\varphi_1{}^\mu = \varphi_3{}^\mu = 0$, and $\varphi_2{}^\mu$, $\varphi_4{}^\mu$ may be taken equal to (1, 0) and (0, 1), respectively. Thus we obtain the four eigenfunctions of the operator S_z:

$$\varphi^1 = \begin{vmatrix}1\\0\\0\\0\end{vmatrix}, \qquad \varphi^2 = \begin{vmatrix}0\\1\\0\\0\end{vmatrix}, \qquad \varphi^3 = \begin{vmatrix}0\\0\\1\\0\end{vmatrix}, \qquad \varphi^4 = \begin{vmatrix}0\\0\\0\\1\end{vmatrix}. \tag{290}$$

We shall now make use of these in order to solve the following problem. Given the ψ, find the probability that an observation of the spin with respect to the z-axis yields the result $+h/4\pi$ (or $-h/4\pi$). Let us apply the procedure of the note in §22 literally; that is, let us develop the matrix ψ in terms of the eigenfunctions φ (the series reduces to four terms; let us indicate by c_λ the coefficients designated by ψ_r^ν in §22). By this method we obtain

$$\psi = c_1\varphi^1 + c_2\varphi^2 + c_3\varphi^3 + c_4\varphi^4.$$

It may be readily verified, by use of (290), that the four coefficients c are identical with ψ_1, ψ_2, ψ_3, ψ_4, respectively. Applying (97′) of §22, we find that the probability corresponding to the eigenvalue s^1 is $P_1 = \iiint |\psi_1|^2\, dx\, dy\, dz$, and similarly the eigenvalue corresponding to s^3 is $P_3 = \iiint |\psi_3|^2\, dx\, dy\, dz$. Since the two eigenvalues coincide, the total probability for the double eigenvalue $+h/4\pi$ is

$$P_+ = \iiint (|\psi_1|^2 + |\psi_3|^2)\, dx\, dy\, dz. \tag{291}$$

Similarly, we should find that the probability for the double eigenvalue $-h/4\pi$ is

$$P_- = \iiint (|\psi_2|^2 + |\psi_4|^2)\, dx\, dy\, dz. \tag{291′}$$

We may therefore say that of the four ψ, the two with odd index correspond (in the vector model) to the spin oriented parallel to the z-axis, and the two with even index to a spin antiparallel to the z-axis. In the nonrelativistic approximation, as we have seen, ψ_3 and ψ_4 may be neglected with respect to ψ_1 and ψ_2, and the formulas (291) and (291′) reduce to those which define the two ψ-functions of Pauli.

54. Plane waves. As an example of the rigorous solution of the Dirac equations, let us examine the special case of an electron not subject to forces and having an exactly determined momentum $\mathbf{p}$. Calling W the total energy (including the rest energy), we shall try to satisfy the equations (269) by a solution which is analogous to the one of the corresponding case in the Schrödinger theory (see §44), namely, by plane waves of frequency W/h and with propagation vector $\mathbf{p}/h$. Let us therefore put

$$\psi_\mu = a_\mu e^{\frac{2\pi i}{h}(-Wt+\mathbf{p}\cdot\mathbf{r})} \qquad (\mu = 1, 2, 3, 4). \tag{292}$$

Substituting into (269), we obtain

$$\left.\begin{aligned}
\left(-\frac{W}{c} + m_0c\right) a_1 + (p_1 - ip_2)a_4 + p_3a_3 &= 0,\\
\left(-\frac{W}{c} + m_0c\right) a_2 + (p_1 + ip_2)a_3 - p_3a_4 &= 0,\\
\left(-\frac{W}{c} - m_0c\right) a_3 + (p_1 - ip_2)a_2 + p_3a_1 &= 0,\\
\left(-\frac{W}{c} - m_0c\right) a_4 + (p_1 + ip_2)a_1 - p_3a_2 &= 0.
\end{aligned}\right\} \tag{293}$$

These four linear homogeneous equations in the four constants a_1, a_2, a_3, a_4 have nonzero solutions only if

$$\begin{vmatrix}
-\frac{W}{c} + m_0c & 0 & p_3 & p_1 - ip_2\\
0 & -\frac{W}{c} + m_0c & p_1 + ip_2 & -p_3\\
p_3 & p_1 - ip_2 & -\frac{W}{c} - m_0c & 0\\
p_1 + ip_2 & -p_3 & 0 & -\frac{W}{c} - m_0c
\end{vmatrix} = 0. \tag{294}$$

Developing the determinant, we find that it is equal to

$$\left(\frac{W^2}{c^2} - m_0^2c^2 - p^2\right)^2,$$

where we have put $p^2 = p_1^2 + p_2^2 + p_3^2$. Hence the condition that the determinant vanish is

$$\frac{W}{c} = \pm\sqrt{m_0^2c^2 + p^2}. \tag{295}$$

This coincides with the relation given by relativistic mechanics between the energy W and the momentum p. If we take the plus sign and expand the radical in a series, we may write, neglecting the higher powers of p,

$$W = m_0c^2 + \frac{p^2}{2m_0} + \cdots \tag{295'}$$

This expression for the energy reduces, for small velocities, to the kinetic energy of ordinary mechanics, increased by the rest energy m_0c^2. If we take the minus sign instead, we obtain a value for W close to $-m_0c^2$; this value has no analogue in ordinary mechanics. We shall return to these anomalous (negative) values of the kinetic energy in §60.

Let us now proceed to a determination of the a. Since, as may be verified, equation (295) makes all third-order minors of the determinant vanish, one may fix two of the four a arbitrarily and solve for the other two. Hence relations (293) will have two linearly independent solutions, corresponding to the two possible orientations of the spin. We shall take for fundamental solutions, in the case $W > 0$,

$$\left.\begin{array}{llll}\text{(I)}\ a_1 = A, & a_2 = 0, & a_3 = A\dfrac{cp_3}{W + m_0c^2}, & a_4 = A\dfrac{c(p_1 + ip_2)}{W + m_0c^2}, \\[2ex] \text{(II)}\ a_1 = 0, & a_2 = A, & a_3 = A\dfrac{c(p_1 - ip_2)}{W + m_0c^2}, & a_4 = A\dfrac{-cp_3}{W + m_0c^2};\end{array}\right\} \tag{296}$$

the modulus of the constant A is determined from the normalization condition (see §10 of Part II).

It is to be noted that the denominators of (296) are of the order of $2m_0c^2$; hence, if $p \ll m_0c$, as it is ordinarily, a_3 and a_4 prove to be small compared with A, as was already seen in general in §51. To the order of approximation in which $|a_3|^2$ and $|a_4|^2$ are considered negligible compared with $|A|^2$, solution (I) corresponds to the case in which the spin is parallel to the z-axis, whereas solution (II) corresponds to the case of antiparallel spin. The most general solution, which is obtained by combining (I) and (II) linearly, represents the cases in which the spin along the z-axis does not have a definite value.

Proceeding now to the case of $W < 0$, we shall find it convenient to take as fundamental solutions

$$\left.\begin{array}{llll} \text{(I)}\ a_1 = B\dfrac{cp_3}{W - m_0c^2}, & a_2 = B\dfrac{c(p_1 + ip_2)}{W - m_0c^2}, & a_3 = B, & a_4 = 0, \\ \text{(II)}\ a_1 = B\dfrac{c(p_1 - ip_2)}{W - m_0c^2}, & a_2 = B\dfrac{-cp_3}{W - m_0c^2}, & a_3 = 0, & a_4 = B. \end{array}\right\} \tag{297}$$

We get $|B|$ from the normalization condition.

In this case a_1 and a_2 are small compared with B (supposing that $p \ll m_0c$). If a_1 and a_2 are considered negligible, solution (I) corresponds to the spin parallel to the z-axis; solution (II), to antiparallel spin.

55. Other form of the Dirac equations. We shall now put the Dirac equation into a different form which, since the variable t is treated symmetrically with x, y, z, lends itself especially to considerations of a relativistic nature.

Let us multiply the Dirac equation (271) by $-i\alpha^4$ (from the left), and let us put ($k = 1, 2, 3$):

$$-i\alpha^4\alpha^k = \gamma^k, \qquad \alpha^4 = \gamma^4, \tag{298}$$

$$P_k = \Pi_k, \qquad iP_4 = \Pi_4. \tag{299}$$

This arrangement permits us to collect even the term with index 4 into a single summation, and the equation may thereby be written in the form[20]

$$\left[\sum_{\mu=1}^{4} \gamma^\mu \Pi_\mu - im_0c\right]\psi = 0. \tag{300}$$

---

[20] We recall that in this whole chapter we indicate by Greek letters the indices taking the values 1, 2, 3, 4, and by ordinary letters those assuming only the values 1, 2, 3.

The matrices γ^μ, defined by (298), are Hermitian, like the α^μ. As may readily be seen, they also satisfy the relations

$$\gamma^\lambda\gamma^\mu + \gamma^\mu\gamma^\lambda = 2\delta_{\lambda\mu} \quad (\lambda, \mu = 1, 2, 3, 4). \tag{301}$$

The operators Π_μ may be written in a more symmetric form by introducing, instead of t, the variable

$$x_4 = ict,$$

and considering A_1, A_2, A_3, iV as the four components of a "four-vector" $\mathbf{\Phi}$ in the space of the variables x_1, x_2, x_3, x_4 (Minkowski space). In fact, it is known from the theory of relativity that under a Lorentz transformation these four quantities transform just like the components of an invariant four-vector. Therefore we put

$$A_k = \Phi_k, \qquad iV = \Phi_4, \tag{302}$$

and then (299) may be assembled into the single expression

$$\Pi_\mu = \frac{h}{2\pi i}\frac{\partial}{\partial x_\mu} + \frac{e}{c}\Phi_\mu. \tag{299'}$$

Let us now deal with the expressions for the (average) electric charge density ρ and the (average) current density $\mathbf{j}$. It is convenient to introduce

$$i\tilde{\psi}\gamma^4 = \psi^+, \tag{303}$$

from which, by multiplying by $-i\gamma^4$ from the right,

$$\tilde{\psi} = -i\psi^+\gamma^4. \tag{303'}$$

The expressions (262) and (264) then become, respectively,

$$\rho = ie\psi^+\gamma^4\psi, \tag{304}$$

$$j_k = -ce\psi^+\gamma^k\psi, \tag{305}$$

which show that the four quantities

$$J_k = j_k, \qquad J_4 = ic\rho \tag{306}$$

(which, we see, also constitute the components of an invariant four-vector, namely, the "four-current") are expressed in a uniform manner by means of the ψ and ψ^+, since (304) and (305) may be combined into the equation

$$J_\mu = -ce\psi^+\gamma^\mu\psi. \tag{307}$$

Note that the equation of continuity (263), in the notation (306), takes on the form

$$\sum_{\mu} \frac{\partial J_\mu}{\partial x_\mu} = 0, \tag{308}$$

which is relativistically invariant.

It may be useful to write down the equation which ψ^+ satisfies. From the Dirac equation (271) we obtain, by taking its complex conjugate (and noting that the conjugate of $\alpha\psi$ is $\tilde{\psi}\alpha$, if α is a Hermitian matrix),

$$\sum_{k} P_k^* \tilde{\psi} \alpha^k - P_4^* \tilde{\psi} + m_0 c \tilde{\psi} \alpha^4 = 0. \tag{309}$$

Inserting (303′) and taking (298) into account, we have

$$\sum_{k} P_k^* \psi^+ \gamma^k + i P_4^* \psi^+ \gamma^4 - i m_0 c \psi^+ = 0. \tag{309′}$$

In order to collect the first two terms under a single summation sign, it is convenient to define the operators

$$\Pi_k^+ = \Pi_k^* = P_k^*, \qquad \Pi_4^+ = -\Pi_4^* = i P_4^*; \tag{310}$$

that is,

$$\Pi_\mu^+ = -\frac{h}{2\pi i} \frac{\partial}{\partial x_\mu} + \frac{e}{c} \Phi_\mu. \tag{310′}$$

Thus equation 309′ becomes

$$\sum_{\mu} \Pi_\mu^+ \psi^+ \gamma^\mu - i m_0 c \psi^+ = 0. \tag{311}$$

56. Relativistic invariance of the Dirac equations. The relativistic invariance of the Dirac equations is to be understood in the following sense. We consider a second system of reference in uniform and rectilinear translation with respect to the first; that is, we pass, by a Lorentz transformation, from the variables x_1, x_2, x_3, $x_4 = ict$, to x_1', x_2', x_3', $x_4' = ict'$. We shall show that in the new reference system, the Dirac equation (300) holds in the same form as in the old frame; that is, it may be written

$$\left[\sum_{\mu=1}^{4} \gamma^\mu \Pi_\mu' - i m_0 c \right] \psi' = 0. \tag{312}$$

This does not mean that the ψ' obtained from this equation is equal to ψ, but that the quantities ρ', $\mathbf{j}'$ (electric charge and current densities in the second system of reference) obtained from ψ' by formulas analogous to (262) and (264) are equal to those obtained from ρ and $\mathbf{j}$ when the latter are transformed according to the laws by which the electric charge and current densities transform in ordinary electromagnetic theory. That is to say, they transform in such a way that the four quantities $J_k = j^k$, $J_4 = ic\rho$ constitute the components of an invariant four-vector in Minkowski space.

In order to prove what we have stated, let us consider the most general Lorentz transformation, or the most general orthogonal transformation in Minkowski space, expressed by the relations

$$x_\mu' = \sum_\nu a_{\mu\nu} x_\nu, \tag{313}$$

where the coefficients $a_{\mu\nu}$ are connected by the relations

$$\sum_\mu a_{\mu\rho} a_{\mu\sigma} = \delta_{\rho\sigma}. \tag{314}$$

It is now to be remembered that these coefficients are real, except for those which contain the index 4 only once; the latter are pure imaginary (such as x_4 and x_4'). The operators transform like the components of a four-vector [as may readily be recognized from (299′)], that is, according to the relations

$$\Pi_\mu' = \sum_\nu a_{\mu\nu} \Pi_\nu. \tag{315}$$

Hence (312) may be written

$$\left[\sum_{\mu,\nu} a_{\mu\nu} \gamma^\mu \Pi_\nu - i m_0 c\right] \psi' = 0. \tag{316}$$

Let us now attempt to satisfy this equation by assuming, pending a posteriori verification, that the ψ transforms linearly, analogous to the behavior of the components of the electric and magnetic fields under a Lorentz transformation. Therefore let us put

$$\psi_\lambda' = \sum_\mu S_{\lambda\mu} \psi_\mu, \tag{317}$$

leaving the coefficients $S_{\lambda\mu}$ undetermined for the present. In short, we shall write, denoting by S the matrix whose elements are $S_{\lambda\mu}$,

$$\psi' = S\psi. \tag{318}$$

Substituting into (316) and multiplying by S^{-1} from the left, we obtain (noting that S commutes with the Π but not with the γ)

$$\left[\sum_{\mu,\nu} a_{\mu\nu} S^{-1}\gamma^{\mu} S \Pi_{\nu} - im_0 c\right]\psi = 0.$$

This equation will be satisfied if the matrix S is such that

$$\sum_{\mu} a_{\mu\nu} S^{-1}\gamma^{\mu} S = \gamma^{\nu}, \tag{319}$$

since in that case the equation is identical with (300), which by hypothesis is satisfied by ψ. Hence we have to prove that for any Lorentz transformation there exists a matrix S satisfying (319). In order to show this, let us first consider an "infinitesimal" Lorentz transformation; that is, let us put

$$a_{\mu\nu} = \delta_{\mu\nu} + \epsilon_{\mu\nu}, \tag{320}$$

where the $\epsilon_{\mu\nu}$ are infinitesimals of the first order. Equation (314) becomes the condition for antisymmetry

$$\epsilon_{\rho\sigma} + \epsilon_{\sigma\rho} = 0. \tag{321}$$

It is obvious in this case to seek for S a form which is infinitely close to the unit matrix, that is, to put $S_{\mu\nu} = \delta_{\mu\nu} + T_{\mu\nu}$, where the $T_{\mu\nu}$ are infinitesimals, or

$$S = \{1\} + T. \tag{322}$$

Then, neglecting infinitesimals of higher order, we have

$$S^{-1} = \{1\} - T, \tag{323}$$

and (319) assumes the form

$$\sum_{\mu} \epsilon_{\mu\nu}\gamma^{\mu} = T\gamma^{\nu} - \gamma^{\nu} T. \tag{324}$$

Now we may easily verify, keeping (301) in mind, that (324) is satisfied by taking

$$T = \tfrac{1}{4}\sum_{\mu,\lambda} \epsilon_{\mu\lambda}\gamma^{\mu}\gamma^{\lambda},$$

or by taking

$$S = \{1\} + \tfrac{1}{4}\sum_{\mu,\lambda} \epsilon_{\mu\lambda}\gamma^\mu\gamma^\lambda. \tag{325}$$

Having thus demonstrated the existence of the matrix S for an infinitesimal transformation, it follows immediately (since, as we know, the Lorentz transformations form a group) that such a matrix exists for any Lorentz transformation and may be constructed as the product of an infinity of matrices of the type (325).

Let us now show that once we have constructed S in this manner, ψ^+ transforms according to the law

$$\psi^{+\prime} = \psi^+ S^{-1}. \tag{326}$$

For this purpose we observe that $\psi^{+\prime}$ is defined by a formula analogous to (303), that is, by

$$\psi^{+\prime} = i\tilde{\psi}'\gamma^4. \tag{327}$$

We may transform (327), noting that we have, from (318),

$$\tilde{\psi}' = \tilde{\psi}\tilde{S}, \tag{328}$$

where we introduce the matrix S defined, as usual, by $\tilde{S}_{\mu\lambda} = S_{\lambda\mu}{}^*$. If we substitute (328) into (327) and keep (303′) in mind, the formula defining $\psi^{+\prime}$ becomes

$$\psi^{+\prime} = \psi^+\gamma^4\tilde{S}\gamma^4. \tag{329}$$

Thus (326) will be proved if we show that S, under any Lorentz transformation, possesses the property

$$\gamma^4\tilde{S}\gamma^4 = S^{-1}, \tag{330}$$

and it will be sufficient to show that this formula holds for an S of the form (325), since we may verify immediately that if it holds for two matrices S_1 and S_2, it will also hold for their product. To prove (330), we note that we obtain from (325):

$$\begin{aligned}\tilde{S} &= \{1\} + \tfrac{1}{4}\sum_{\mu,\lambda} \epsilon_{\mu\lambda}{}^*\widetilde{\gamma^\mu\gamma^\lambda}\\ &= \{1\} + \tfrac{1}{4}\sum_{h,k} \epsilon^*_{hk}\widetilde{\gamma^h\gamma^k} + \tfrac{1}{4}\sum_{k} (\epsilon^*_{4k}\widetilde{\gamma^4\gamma^k} + \epsilon^*_{k4}\widetilde{\gamma^k\gamma^4}).\end{aligned}$$

From this, recalling that for Hermitian matrices such as γ one has

$\widetilde{AB} = \tilde{B}\tilde{A} = BA$, and that in addition $\epsilon_{hk}^* = \epsilon_{hk}$, $\epsilon_{4k}^* = -\epsilon_{4k}$, $\epsilon_{k4}^* = -\epsilon_{k4}$, we obtain

$$\tilde{S} = \{1\} - \tfrac{1}{4}\sum_{h,k} \epsilon_{hk}\gamma^k\gamma^h - \tfrac{1}{4}\sum_k \epsilon_{4k}(\gamma^k\gamma^4 - \gamma^4\gamma^k).$$

Multiplying both sides by γ^4 from the right and from the left, and recalling (301), we find that

$$\begin{aligned}\gamma^4\tilde{S}\gamma^4 &= \{1\} - \tfrac{1}{4}\sum_{h,k} \epsilon_{hk}\gamma^k\gamma^h - \tfrac{1}{4}\sum_k \epsilon_{4k}(\gamma^4\gamma^k - \gamma^k\gamma^4)\\ &= \{1\} - \tfrac{1}{4}\sum_{\mu,\lambda} \epsilon_{\mu\lambda}\gamma^\mu\gamma^\lambda = \{1\} - T = S^{-1};\end{aligned}$$

thus (330) is proved.

Having found the transformation law[21] for ψ and ψ^+, we may find it immediately for the components of the four-current, that is, of the J_μ defined by (307). In fact, we have

$$J_\mu' = -ce\psi^+S^{-1}\gamma^\mu S\psi^+,$$

and since we find from (319) that

$$S^{-1}\gamma^\mu S = \sum_\nu a_{\mu\nu}\gamma^\nu,$$

we have

$$J_\mu' = \sum_\nu a_{\mu\nu}J_\nu, \tag{331}$$

which proves that the J_μ transform like the components of an invariant four-vector; this completes the proof.

57. Gauge invariance of the Dirac equations. In the Dirac equations the electromagnetic field is represented by the potentials V and **A**. Now it is well known that these potentials are not physically determined in a unique manner, since we may also attribute to the same electromagnetic field (as may readily be

---

[21] A group of four quantities which transform like ψ_1, ψ_2, ψ_3, ψ_4 in a Lorentz transformation is called a *spinor* (concept analogous to that of *tensor*). This concept of van der Waerden and others constitutes the basis of a systematic treatment analogous to tensor calculus (see *Göttinger Nachr.* 1929, page 100). The components of a spinor do not stand for directly observable physical quantities (which would be contrary to the principle of relativity) but are of interest inasmuch as they determine indirectly the values of observable quantities.

shown) the potentials

$$V' = V - \frac{1}{c}\frac{\partial \varphi}{\partial t}, \qquad \mathbf{A}' = \mathbf{A} + \operatorname{grad} \varphi, \tag{332}$$

where $\varphi(x, y, z, t)$ is an arbitrary function. If we use the four-potential Φ defined by (302), we can say that the Φ_μ may be replaced by $\Phi_\mu' = \Phi_\mu + \partial\varphi/\partial x_\mu$. But since this substitution does not correspond to any physical change, it is necessary that the Dirac equations be insensitive to this alteration. More precisely, none of the observable results furnished by the equations must be altered by this change, such as for instance the average values of the electric charge and current densities, that is, the J_μ. (On the other hand, ψ may change, since it has no direct physical meaning.) This property, which we shall now verify, is called *gauge invariance* (*Eichinvarianz* in German).

Let us write the Dirac equation in the form (300), replacing the Φ_μ by the Φ_μ', and ψ by a new function ψ'. We have, writing the operators explicitly by means of (299′),

$$\left[\sum_\mu \gamma^\mu \left(\frac{h}{2\pi i}\frac{\partial}{\partial x_\mu} + \frac{e}{c}\Phi_\mu + \frac{e}{c}\frac{\partial \varphi}{\partial x_\mu}\right) - im_0c\right]\psi' = 0.$$

Now, this equation may be satisfied by taking

$$\psi' = \psi\, e^{-\frac{2\pi i}{h}\frac{e}{c}\varphi}, \tag{333}$$

since this substitution brings the equation back to (300), as may easily be verified. To this ψ' there corresponds, by virtue of (303),

$$\psi'^{+} = i\psi'\gamma^4 = \psi^{+}\, e^{\frac{2\pi i}{h}\frac{e}{c}\varphi};$$

and, upon forming with these expressions for ψ' and ψ'^{+} the J'_μ by formula (307), we see that the exponentials cancel and the arbitrary function φ disappears.

To make the statement more general: according to the fundamental principle of quantum mechanics (§22), the probability of a given result G_n in the measurement of an observable G, when the system is in the state ψ, is given, as we have seen, by $|\psi \cdot \varphi_n|^2$, where φ_n does not depend on ψ but only on the observable G and its eigenvalue G_n. If ψ is replaced by ψ', the product $\psi \cdot \varphi_n$ will change only by a factor of modulus 1, and hence the square of its modulus is unchanged. Thus the desired probability turns out to be the same.

58. The electron in a central field. As an application of the Dirac theory, we shall now deal with the problem of an electron subject only to the action of an electrostatic central force of any type. The relations found in this way will then be used in the following section (with the law of force specialized to the Coulomb type) to treat the problem of hydrogenlike systems. It will be noted that the results obtained correspond qualitatively to those found by the modelistic spin theory: that is, by using the Bohr-Sommerfeld model, improved by the introduction of the magnetic moment of the electron, considered as an ordinary magnetic dipole but subject to the spatial quantization rule (see §62 of Part II). The Dirac theory, however, in addition to being rationally more coherent, yields results which are slightly different and in better agreement with experience.

Denoting, as usual, by $U = -eV(r)$ the potential energy of the central field in which the electron is located, let us consider a stationary state of energy W. The Dirac equations (272) will assume the form

$$\left.\begin{aligned}
\frac{2\pi i}{h}\frac{(-W+U+m_0c^2)}{c}\psi_1 + \left(\frac{\partial}{\partial x_1} - i\frac{\partial}{\partial x_2}\right)\psi_4 + \frac{\partial}{\partial x_3}\psi_3 &= 0,\\
\frac{2\pi i}{h}\frac{(-W+U+m_0c^2)}{c}\psi_2 + \left(\frac{\partial}{\partial x_1} + i\frac{\partial}{\partial x_2}\right)\psi_3 - \frac{\partial}{\partial x_3}\psi_4 &= 0,\\
\frac{2\pi i}{h}\frac{(-W+U-m_0c^2)}{c}\psi_3 + \left(\frac{\partial}{\partial x_1} - i\frac{\partial}{\partial x_2}\right)\psi_2 + \frac{\partial}{\partial x_3}\psi_1 &= 0,\\
\frac{2\pi i}{h}\frac{(-W+U-m_0c^2)}{c}\psi_4 + \left(\frac{\partial}{\partial x_1} + i\frac{\partial}{\partial x_2}\right)\psi_1 - \frac{\partial}{\partial x_3}\psi_2 &= 0.
\end{aligned}\right\} \qquad (334)$$

As we have seen in §46 of Part II, the Schrödinger equation corresponding to this problem has solutions of the type $\psi = Y_{lm}f(r)$, where Y_{lm} is a surface harmonic whose indices l and m represent the azimuthal quantum number and the magnetic quantum number of the electron, respectively. Let us try to satisfy equations (334) analogously by taking each of the four ψ of the form

$$\psi_\mu = Z(\theta, \varphi)f_\mu(r). \qquad (335)$$

The spherical harmonic is here denoted[22] by Z rather than by Y_{lm} because it is convenient to normalize it (following Darwin's treatment[23]) in a way different from that in §46 of Part II, in order to simplify the expressions. To be exact, we shall take

$$Z_l^m = \sqrt{\frac{4\pi}{2l+1}(l+m)!(l-m)!}\, Y_{lm}. \tag{336}$$

It may then be verified, using the transformation relations for passing from the Cartesian coordinates x_1, x_2, x_3 to polar coordinates r, θ, φ, that for the derivatives of expressions of the type fZ_l^m (where $f = f(r)$ is any function of r) the following relations hold:

$$\left.\begin{aligned}
\left(\frac{\partial}{\partial x_1} + i\frac{\partial}{\partial x_2}\right) fZ_l^m &= \frac{1}{2l+1}\left[\left(\frac{df}{dr} - \frac{l}{r}f\right) Z_{l+1}^{m+1}\right. \\
&\qquad \left. - (l-m)(l-m-1)\left(\frac{df}{dr} + \frac{l+1}{r}f\right) Z_{l-1}^{m+1}\right], \\
\left(\frac{\partial}{\partial x_1} - i\frac{\partial}{\partial x_2}\right) fZ_l^m &= \frac{1}{2l+1}\left[-\left(\frac{df}{dr} - \frac{l}{r}f\right) Z_{l+1}^{m-1}\right. \\
&\qquad \left. + (l+m)(l+m-1)\left(\frac{df}{dr} + \frac{l+1}{r}f\right) Z_{l-1}^{m-1}\right], \\
\frac{\partial}{\partial x_3} fZ_l^m &= \frac{1}{2l+1}\left[\left(\frac{df}{dr} - \frac{l}{r}f\right) Z_{l+1}^{m}\right. \\
&\qquad \left. + (l+m)(l-m)\left(\frac{df}{dr} + \frac{l+1}{r}f\right) Z_{l-1}^{m}\right].
\end{aligned}\right\} \tag{337}$$

With the aid of these formulas we may show without difficulty that if we take the four ψ of the form

$$\left.\begin{aligned}
\psi_1 &= ia_1F_+(r)Z_{l+1}^m, & \psi_2 &= ia_2F_+(r)Z_{l+1}^{m+1}, \\
\psi_3 &= a_3G_+(r)Z_l^m, & \psi_4 &= a_4G_+(r)Z_l^{m+1}
\end{aligned}\right\} \tag{338}$$

(where the a are constants and F_+, G_+ are two functions of r, undetermined for the present), and if we substitute them into (334), it then suffices to impose upon the constants a the conditions

$$\frac{a_2}{a_1} = 1, \qquad \frac{a_3}{a_4} = -\frac{l+m+1}{l-m}$$

for the spherical harmonics to be eliminated from the equations.

---

[22] For simplicity of notation, we shall omit writing the indices of the spherical harmonic Z. These will generally be different for the four ψ, as we shall see later on.

[23] Proc. Roy. Soc. **A118**, 554 (1928).

Then the equations reduce to only two (since the first and second equations become identical, and likewise the third and fourth), as follows:

$$\left.\begin{aligned} -\frac{2\pi}{h}\frac{(-W+U+m_0c^2)}{c}a_1F_+ - \frac{a_4}{l-m}\left(\frac{dG_+}{dr}-\frac{l}{r}G_+\right) &= 0,\\ -\frac{2\pi}{h}\frac{(-W+U-m_0c^2)}{c}\frac{a_4}{l-m}G_+ + a_1\left(\frac{dF_+}{dr}+\frac{l+2}{r}F_+\right) &= 0.\end{aligned}\right\}$$

The constants a_1 and a_4 remain arbitrary, and we shall take them equal to 1 and to $-(l-m)$, respectively, so that we have

$$a_1 = 1, \qquad a_2 = 1, \qquad a_3 = l+m+1, \qquad a_4 = -(l-m), \tag{339}$$

and the preceding equations become

$$\left.\begin{aligned} -\frac{2\pi}{h}\frac{(-W+U+m_0c^2)}{c}F_+ + \frac{dG_+}{dr} - \frac{l}{r}G_+ &= 0,\\ \frac{2\pi}{h}\frac{(-W+U-m_0c^2)}{c}G_+ + \frac{dF_+}{dr} + \frac{l+2}{r}F_+ &= 0.\end{aligned}\right\} \tag{340}$$

It is to be noted that a solution of the form considered here can exist only if m lies between $-(l+1)$ and l (endpoints included); otherwise there would occur spherical harmonics whose upper index would be larger (in absolute value) than the lower one, and we have not assigned any meaning to such symbols. As far as l is concerned, it may take the values 0, 1, 2, ...

A second way of satisfying (334) consists in taking the ψ of the form

$$\left.\begin{aligned} \psi_1 &= ib_1F_-(r)Z_{l-1}^{m}, & \psi_2 &= ib_2F_-(r)Z_{l-1}^{m+1},\\ \psi_3 &= b_3G_-(r)Z_l^m, & \psi_4 &= b_4G_-(r)Z_l^{m+1}.\end{aligned}\right\} \tag{341}$$

Substituting into (334) and proceeding as above, we find that if we take

$$b_1 = (l+m), \qquad b_2 = -(l-m-1), \qquad b_3 = 1, \qquad b_4 = 1, \tag{342}$$

these equations reduce to the two following relations in the functions $F(r)$ and $G(r)$:

$$\left.\begin{aligned} -\frac{2\pi}{h}\frac{(-W+U+m_0c^2)}{c}F_- + \frac{dG_-}{dr} + \frac{l+1}{r}G_- &= 0\\ \frac{2\pi}{h}\frac{(-W+U-m_0c^2)}{c}G_- + \frac{dF_-}{dr} - \frac{l-1}{r}F_- &= 0.\end{aligned}\right\} \tag{343}$$

Note that these equations differ from (340), which are satisfied by F_+ and G_+, simply by the substitution of $-(l+1)$ for l. A solution of this form can exist only if m lies between $-l$ and $l-1$, endpoints included. Furthermore, there is no solution for $l = 0$, since there are no spherical harmonics with negative lower index.

To interpret the two types of solutions found in this way, we recall (see §53) that the total angular momentum with respect to the z-axis corresponds to the operator

$$\mathfrak{N}_z = \mathfrak{M}_z + \mathfrak{S}_z = \frac{h}{2\pi i}\frac{\partial}{\partial\varphi} + \frac{h}{4\pi}\begin{vmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{vmatrix}.$$

Let us now apply this operator to the ψ of the form (338) or (341), noting that

$$\frac{\partial}{\partial\varphi} Z_l^m = imZ_l^m.$$

We find in both cases that

$$\mathfrak{N}_z\psi = \left(m + \frac{1}{2}\right)\frac{h}{2\pi}\psi,$$

which shows that all the solutions which we have here considered represent states in which the angular momentum with respect to the z-axis has a definite value, specifically the value $(m + \frac{1}{2})h/2\pi$. Keeping in mind the limits between which m may vary, we see that N_z can vary, in case (338), from $-(l+\frac{1}{2})h/2\pi$ to $+(l+\frac{1}{2})h/2\pi$ and, in case (341), from $-(l-\frac{1}{2})h/2\pi$ to $+(l-\frac{1}{2})h/2\pi$. These results are identical with those obtained from the modelistic theory, if N_z is interpreted as the projection of the total angular momentum upon the z-axis and if the total angular momentum is taken as equal to $jh/2\pi$, where j is the *inner quantum number* (see §62 of Part II); that is, $j = l + \frac{1}{2}$ for case (338), and $j = l - \frac{1}{2}$ for case (341). The first solution therefore corresponds, in the modelistic interpretation, to the case of the spin parallel to the orbital angular momentum, and the second solution to the spin antiparallel to that momentum.

Finally, in the case where $l = 0$, we have pointed out that the solution (341) is absent, which means that N_z can have only the value $\frac{1}{2}\,h/2\pi$ (or j only the value $\frac{1}{2}$), which also follows immediately from the vector model.

59. Theory of hydrogenlike systems (fine-structure). We shall now apply the results of the preceding section to the case of a hydrogenlike system; that is, we shall specialize the function U by taking

$$U = -\frac{Zc^2}{r}. \tag{344}$$

First let us treat the case of solution (338), that is, of $j = l + \frac{1}{2}$, and let us seek the eigenvalues (for the parameter W) and the eigenfunctions of equations (340), limiting ourselves to the case in which the energy $E = W - m_0c^2$ is negative (case corresponding to elliptical orbits). We introduce the notation

$$A^2 = \frac{2\pi}{h}\left(m_0c + \frac{W}{c}\right), \qquad B^2 = \frac{2\pi}{h}\left(m_0c - \frac{W}{c}\right), \tag{345}$$

and in addition we introduce the fine-structure constant $\alpha = 2\pi\, e^2/hc$. Equation (340) may then be written

$$\left.\begin{aligned} \left(-B^2 + \frac{\alpha Z}{r}\right)F_+ + \frac{dG_+}{dr} - \frac{l}{r}G_+ &= 0, \\ \left(-A^2 - \frac{\alpha Z}{r}\right)G_+ + \frac{dF_+}{dr} + \frac{l+2}{r}F_+ &= 0. \end{aligned}\right\} \tag{346}$$

We observe that for r approaching infinity, these equations approach the form

$$-B^2F_+ + \frac{dG_+}{dr} = 0, \qquad -A^2G_+ + \frac{dF_+}{dr} = 0,$$

from which we obtain

$$\frac{d^2F_+}{dr^2} - A^2B^2F_+ = 0, \qquad \frac{d^2G_+}{dr^2} - A^2B^2G_+ = 0.$$

Hence the solutions F_+ and G_+ of (346) will have $e^{\pm ABr}$ for their asymptotic expression. Since we are looking for solutions going to zero for $r \rightarrow \infty$, we must discard the plus sign. Thus we are led to look for solutions of the form

$$\left.\begin{aligned} F_+ &= e^{-ABr}(a_0r^\gamma + a_1r^{\gamma+1} + \cdots), \\ G_+ &= e^{-ABr}(b_0r^\gamma + b_1r^{\gamma+1} + \cdots). \end{aligned}\right\} \tag{347}$$

The other possible singular point is $r = 0$. There the preceding expressions are regular if $\gamma \geqslant 0$; whereas if $\gamma < 0$, they become infinite to the order r^γ. Such a singularity is acceptable, provided that $|\gamma| < 1$ (see §28 of Part II).

Substituting these expressions into (346) and setting the coefficients of $r^{\gamma-1}$ equal to zero at the same time, we find

$$\left.\begin{aligned} Z\alpha a_0 + (\gamma - l)b_0 &= 0, \\ (\gamma + l + 2)a_0 - Z\alpha b_0 &= 0; \end{aligned}\right\}$$

and upon setting the determinant of the coefficients of these two linear equations in a_0, b_0 equal to zero, we find for γ the equation

$$(\gamma - l)(\gamma + l + 2) + Z^2\alpha^2 = 0,$$

which yields (discarding the solution with the minus sign, which would give $\gamma < -1$)

$$\gamma = -1 + \sqrt{(l + 1)^2 - Z^2\alpha^2}. \tag{348}$$

Now upon setting the coefficient of $r^{\gamma+s}$ ($s = 0, 1, 2, \ldots$) equal to zero, we find

$$\left.\begin{aligned} -B(Ba_s - Ab_s) + Z\alpha a_{s+1} + (\gamma + s - l + 1)b_{s+1} &= 0, \\ A(Ba_s - Ab_s) - Z\alpha b_{s+1} + (\gamma + s + l + 3)a_{s+1} &= 0, \end{aligned}\right\} \tag{349}$$

from which we obtain, multiplying the first by A and the second by B and adding,

$$a_{s+1}[AZ\alpha + B(\gamma + s + l + 3)] + b_{s+1}[A(\gamma + s - l + 1) - BZ\alpha] = 0.$$

It is now convenient to introduce a single constant c_{s+1} in place of a_{s+1} and b_{s+1}, as follows:

$$\begin{aligned} a_{s+1} &= c_{s+1}[A(\gamma + s - l + 1) - BZ\alpha], \\ b_{s+1} &= -c_{s+1}[AZ\alpha + B(\gamma + s + l + 3)]. \end{aligned}$$

Substituting into (349) these expressions and analogous ones for a_s and b_s, we find the recursion formula for the c_s:

$$c_{s+1} = \frac{-Z\alpha(A^2 - B^2) - 2AB(\gamma + s + 1)}{Z\alpha^2 + (\gamma + s + l + 3)(\gamma + s - l + 1)}\, c_s. \tag{350}$$

The functions (347) will certainly be zero at infinity if the series reduce to polynomials. Calling n' the degree of the latter, we must have for this purpose $c_{n'} \neq 0$ and $c_{n'+1} = 0$, and hence, because of (350),

$$\frac{Z\alpha(B^2 - A^2)}{2AB} = \gamma + n' + 1,$$

or, by (348),

$$Z\alpha \frac{B^2 - A^2}{2AB} = \sqrt{l(l+1)^2 - Z^2\alpha^2} + n'.$$

Substituting expressions (345) for A and B and solving for W, we find

$$W = m_0c^2\left[1 + \frac{Z^2\alpha^2}{(\sqrt{(l+1)^2 - Z^2\alpha^2} + n')^2}\right]^{-\frac{1}{2}}; \tag{351}$$

and recalling that for the solution with which we are dealing here we have $j = l + \frac{1}{2}$, we obtain

$$W = m_0c^2\left[1 + \frac{Z^2\alpha^2}{(\sqrt{(j+\frac{1}{2})^2 - Z^2\alpha^2} + n')^2}\right]^{-\frac{1}{2}}. \tag{352}$$

If we proceed now to consider the solution (341) corresponding to $j = l - \frac{1}{2}$, it is not necessary to repeat the calculation, since it is sufficient to note that equations (343) differ from equations (340) only by the substitution of $-(l + 1)$ for l. Hence upon performing this same substitution in formula (351), we find

$$W = m_0c^2\left[1 + \frac{Z^2\alpha^2}{(\sqrt{l^2 - Z^2\alpha^2} + n')^2}\right]^{-\frac{1}{2}};$$

but then introducing $j + \frac{1}{2}$ instead of l, we are back to formula (352). The latter therefore expresses the energy levels in both cases. Let us develop it in a power series in α up to terms in α^4 inclusive (an approximation which is more than sufficient for a comparison with experiment), and let us remove the term m_0c^2 which represents the rest energy. We then find for the energy E (other than the rest energy) the following expression, where we have set $n = j + \frac{1}{2} + n'$ and where we have used the fact that $m_0c^2\alpha^2 = 2hcR$ (R = Rydberg constant):

$$E_{n'} = -\frac{Z^2Rhc}{n^2}\left[1 - \frac{Z^2\alpha^2}{n^2}\left(\frac{3}{4} - \frac{n}{j+\frac{1}{2}}\right)\right]. \tag{353}$$

If we neglect the term in α^2, we find again the well-known expression for the Balmer terms as given by the Schrödinger equation (see §48 of Part II). Thus we recognize that the n which was just introduced (and which is an integer) must be identified with the principal quantum number. If, however, we take the correction term in α^2 into account, we see that upon fixing the value of n, we obtain from this formula, instead of the single energy level of the Schrödinger theory, n different energy levels, corresponding to the n values which may be taken by j (from $\frac{1}{2}$ to $n - \frac{1}{2}$). These levels, however, lie very close together, since j occurs only in the term in α^2. This splitting is due to the combined effects of the relativistic correction and the spin, which are both contained intrinsically in the Dirac equations. To the splitting of the levels there naturally corresponds a splitting of the spectral lines, which constitutes their *fine-structure* (see §60 of Part II). The line emitted in the transition from a level of principal quantum number n to one of principal quantum number n' must be made up of nn' components, of which some, however, are excluded by the selection rules of the inner quantum number j (see §64d of Part II).

The formula just found for the fine-structure of hydrogenlike spectra differs from (341) in §60 of Part II (which resulted from the application of the relativistic correction to the Bohr-Sommerfeld theory) only in having the number $j + \frac{1}{2}$ in place of k. But since both $j + \frac{1}{2}$ and k take on the same series of values (from 1 to n), the two formulas lead to exactly the same energy levels (although associated with different quantum numbers) and hence to the same line structure, which agrees quite well with the one which is observed experimentally. However, the results of the two theories are different with regard to the intensity, which may be calculated, in the case of the old theory, by means of the correspondence principle, and in the new theory by means of the Dirac theory of radiation. We shall omit these calculations,[24] confining ourselves to stating that experience is in favor of the Dirac theory. For a detailed comparison between the theory outlined above and the observed fine-structure of the spectra of hydrogen and He$^+$ (which latter lends itself better to an experimental investigation, as was already pointed out), we refer to No. 1 of the Bibliography, page 316. We restrict ourselves here to reproducing, as an example, a graphical

[24] See, for instance, No. 1 of the Bibliography, page 447.

comparison for the line λ4686 of He$^+$, corresponding to the transition from $n = 4$ to $n = 3$ (Fig. 47). The curve represents the intensity as a function of the frequency, as obtained from spectrophotometric measurements of the line. The vertical line segments indicate the positions of the components as predicted by the theory, their length being a measure of the calculated intensity, in arbitrary units (the comparison concerns only the *ratios* of the intensities of the various components, and the distances between them).

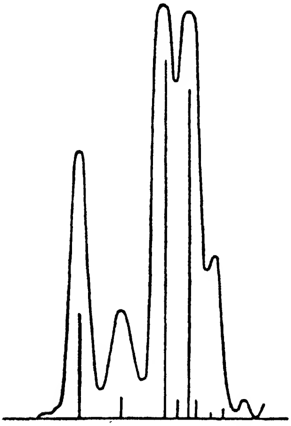

Fig. 47

60. States of negative kinetic energy and positron theory. The Dirac equation possesses (as we have seen in §54 for the particular case of plane waves and $U = 0$), alongside any solution representing a stationary state with positive kinetic energy $\mathfrak{I}$,[25] an analogous solution with negative kinetic energy. This arises from the fact that in relativistic mechanics, the kinetic energy $\mathfrak{I}$ of a particle is related to its momentum by the equation

$$\frac{\mathfrak{I}^2}{c^2} = m_0^2c^2 + p^2,$$

which is quadratic in $\mathfrak{I}$, whereas the corresponding relation of classical mechanics is linear. Therefore even in classical (non-quantized) relativistic mechanics there is a possibility of having solutions with negative kinetic energy. Specifically, we obtain from the last equation (neglecting to write down the terms of order higher than the second in v/c)

$$\mathfrak{I} = \pm\left(m_0c^2 + \frac{p^2}{2m_0} + \cdots\right).$$

We see that $\mathfrak{I}$ may vary over two intervals (from $-\infty$ to $-m_0c^2$ and from m_0c^2 to $+\infty$) which are separated from each other.

---

[25] For short, we shall also include the rest energy m_0c^2 in the kinetic energy, and shall designate it by $\mathfrak{I}$ in order to distinguish it from the ordinary kinetic energy T. We have $\mathfrak{I} = m_0c^2 + T$. The total energy $\mathfrak{I} + U$ will then be denoted by W as usual. In the case of §54 it was assumed that $U = 0$, and hence $W = \mathfrak{I}$.

Therefore, since in *non*-quantized mechanics the kinetic energy can only vary continuously, it may not pass from one interval to the other and hence, if the motion is started with positive $\mathfrak{I}$, states with negative kinetic energy no longer have to be considered. In quantum mechanics, however, these considerations do not hold, because we may no longer assert that $\mathfrak{I}$ varies continuously. There are cases now in which there is a finite probability, or even the certainty, that the electron passes from a state with positive $\mathfrak{I}(> m_0c^2)$ to one with negative $\mathfrak{I}(< -m_0c^2)$. O. Klein has pointed out a typical case,[26] in which an electron with total energy $W(> 0)$ is directed at a potential step such as the one in Fig. 25, of height $U_0 > W + m_0c^2$. Then, if ψ to the left of the step is represented by (292), it is represented by analogous waves on the right, in which $\mathbf{p}$ is replaced by a real $\mathbf{p}'$ such that

$$\frac{W - U_0}{c} = -\sqrt{m_0^2c^2 + p'^2}.$$

Hence the waves continue past the step (in sinusoidal form, not in exponential form as in the Schrödinger theory; see §40 of Part II), where the kinetic energy is $\mathfrak{I} = W - U_0$ and is therefore negative. This means that the electron has a certain probability, which may be considerable, of overcoming the barrier by simultaneously passing into a state with negative kinetic energy.[27] As a matter of fact, even without the intervention of external forces like the ones considered by Klein, transitions to states of negative kinetic energy may occur. From the radiation theory of Dirac we even get the result that as soon as an electron is put into a state of positive kinetic energy, it would undergo such a transition immediately and spontaneously, while radiating the energy difference.

The properties of an electron with negative kinetic energy should be rather singular: when placed in an electric and magnetic field, it would acquire an acceleration directed in the sense opposite to that of an ordinary electron (and would behave in this respect like a

---

[26] *Zeits. f. Physik* **53**, 157 (1929).

[27] The analytical reason for this fact is that the plane waves with positive kinetic energy do not by themselves constitute a *complete* system of orthogonal functions. In order to have a complete set, we must add the waves with negative kinetic energy. Hence, given an arbitrary initial ψ, it is not in general possible to expand it in Fourier integrals without also employing terms with negative kinetic energy.

positive electron); what is even more unusual, its energy would continue to decrease as its velocity increased.

In order to show that a negative electron of kinetic energy $\mathfrak{T}(< 0)$ moves as would a positive electron with positive kinetic energy $-\mathfrak{T}$ in the same field, let us denote by ψ an eigenfunction (matrix) of the former, relative to a stationary state of negative kinetic energy $\mathfrak{T}$; it will satisfy the Dirac equation

$$(\Sigma\alpha^k P_k - P_4 + m_0 c\alpha^4)\psi = 0, \tag{354}$$

with

$$P_k = \frac{h}{2\pi i}\frac{\partial}{\partial x_k} + \frac{e}{c}A_k, \qquad P_4 = \frac{1}{c}(W + eV) = \frac{\mathfrak{T}}{c}. \tag{355}$$

On the other hand, a positive electron of kinetic energy $-\mathfrak{T}$ would correspond to an eigenfunction ψ' satisfying the equation

$$(-\Sigma\alpha^k P_k^* + P_4^* + m_0 c\alpha^4)\psi' = 0, \tag{356}$$

obtained from the preceding one by changing e into $-e$ and $\mathfrak{T}$ into $-\mathfrak{T}$.

We shall now show that (356) is satisfied by taking

$$\psi' = S\psi^*, \qquad \text{with} \qquad S = \alpha^4\alpha^2. \tag{357}$$

In fact, the matrix S defined in this way has the following property:

$$\alpha^k S = S\alpha^{k*}, \qquad \alpha^4 S = -S\alpha^{4*}, \tag{358}$$

which may be verified immediately by applying (266) and noting that because of (267), the complex conjugates of the matrices α are equal to the same matrices, except for α^{2*}, which is equal to $-\alpha^2$. Hence if we substitute (357) into (356) and take (358) into account, the left member becomes

$$-S(\Sigma\alpha^{k*}P_k^* - P_4^* + m_0 c\alpha^{4*})\psi^*;$$

that is, it is equal to the complex conjugate of the left member of (354), multiplied by $-S$ from the left. Therefore the last expression is zero by virtue of (354), and (356) is satisfied. Since (357) holds for any eigenfunction, it will also be valid for a sum of eigenfunctions; so that if ψ represents a wave packet corresponding to a negative electron with negative kinetic energy (which means that ψ is formed, at least *predominantly*, with eigenfunctions corresponding to negative values of $\mathfrak{I}$), then $\psi' = \alpha^4\alpha^2\psi^*$ represents a wave packet corresponding to a positive electron with positive kinetic energy. And since, as we may readily verify, $|\psi'|^2 = |\psi|^2$, the two packets coincide and move in the same manner.

It is not legitimate, however, to go ahead and identify a negative electron with negative kinetic energy with a positron. It suffices to observe that its kinetic energy as a function of the momentum p is expressed in first approximation by

$$\mathfrak{I} = -m_0c^2 - \frac{1}{2m_0}p^2, \tag{359}$$

and hence decreases with increasing momentum or increasing velocity, which certainly does not correspond to the properties of any of the experimentally observed particles.

From the very beginning, the existence of negative kinetic energy states and the impossibility of ignoring them seemed to constitute one of the most serious objections to the Dirac theory. In order to obviate this difficulty, Dirac proposed[28] an ingenious interpretation of these states, which led to the prediction of the existence of a positive electron and showed the way to calculate its properties. This theory, although still not devoid of weak and obscure points, has earned great acclaim, especially since the experimental discovery of the positron. Here, in brief, is what is involved.

Dirac makes the hypothesis that a space devoid of electric charges must not be thought of as a space in which there are no electrons; rather, it must be considered to contain an infinite number of electrons with negative energy, and exactly one for each stationary state. This particular distribution of electrons, which we shall call "normal," is stable by virtue of the Pauli principle, which prevents an electron from occupying a state which is already occupied by another electron. The "normal" distribution is not to give rise to any electric field; for this reason it is necessary to make the further hypothesis that in the equation div $E = 4\pi\rho$ we are always to understand by ρ, not the charge due to all electrons present (which would be infinite) but only the difference between that charge and the charge due to the "normal" distribution. If we now modify the normal distribution, by raising one of the electrons

---

[28] *Proc. Roy. Soc.* **A126,** 360 (1931); *ibid.,* **A133,** 60 (1931). See also No. 6 of the Bibliography.

from one of the negative energy states (represented, for instance, by the eigenfunction $\psi_{n'}$) to a state of positive energy (whose eigenfunction is to be $\psi_{n''}$), the newly obtained distribution will differ from a normal one in two respects: the presence of an electron in the state n'', and the absence of an electron from the state n'. The first of these facts gives rise to all the ordinary manifestations of an ordinary electron in the state n'', and the "hole" left in the negative energy states will manifest itself as an electron with positive charge, or *positron*, in the state n'. More generally, a positron in a general state will be represented by *subtracting* from the normal distribution a term $\sum_n c_n\psi_n$ (the sum being extended over all or part of the negative energy states), that is, by subtracting from the normal distribution a wave packet with negative kinetic energy, which, as we have seen, moves in the same way as an electron of charge $+e$. Thus, interpreting a positron not as an *existing* wave packet but as a *missing* wave packet, we eliminate the above-mentioned difficulty concerning the energy. In fact, if the velocity of the positrons increases, the missing kinetic energy of the system decreases, or, in other words, its total kinetic energy increases, as it should.[29]

As an obvious corollary of this theory, it should be possible to create an electron-positron pair. To a certain extent this phenomenon would be analogous to the ordinary excitation of an atom, since it would consist in bringing an electron from a state of negative kinetic energy (where the latter does not reveal itself in any way) to a state of positive kinetic energy. It is to be noted, however, since the energy of the first state is below $-m_0c^2$ and that of the final state is above m_0c^2, that the energy necessary for pair-creation should ever so slightly exceed $2m_0c^2$, which amounts to about one million volts. This energy might be furnished by X rays or gamma rays, provided they are of a frequency ν such that $h\nu > 2m_0c^2$ (that is, $\tilde{\nu} > 8.3 \times 10^9$ cm^{-1}). In that case the photon may "materialize"; that is, it may transform into a pair of electrons,

---

[29] In the paper cited, Dirac interpreted the "holes" as protons, since the existence of the positron was then unknown. However, it was not explained why the mass of the positive particles was so different from the mass of the electron. With the experimental discovery of the positron (1932), the Dirac "holes" found their most direct and natural interpretation.

one positive and one negative, having between them a kinetic energy $h\nu - 2m_0c^2$. One may further calculate, on the basis of the Dirac theory, the probability for this *materialization of radiation* (*pair-creation*) to occur, and it is found that the phenomenon is possible in practice only when the photon traverses the electric field of a nucleus.

These theoretical predictions are in good agreement with numerous experimental results, principally those of I. Curie and F. Joliot, Chadwick, Blackett and Occhialini, Anderson and Neddermeyer, and Meitner and Philipp.[30] From these experiments it turns out, for instance, that if the gamma rays of thorium C″ (whose photons have an energy of 2.65×10^6 volts) are made to impinge upon a lead screen in a Wilson chamber, positive and negative electrons will leave the lead (the latter being considerably more abundant) whose sign and velocity may be determined by deflecting them in a magnetic field. It is found that whereas the kinetic energy of the negative electrons may reach a value equal to the entire photon energy, the kinetic energy of the positive electrons is limited with sufficient sharpness to not more than 1.6×10^6 volts. This result is interpreted as follows: the negative electrons are due, in addition to the materialization of the photons, to the ordinary photoelectric effect (in which, as we know, almost the entire energy of the photon may transform into kinetic energy), whereas the positrons are due to pair-creation exclusively. Since this phenomenon absorbs an energy of about one million volts, the pair which is formed will have a kinetic energy of $2.6 - 1 = 1.6$ million volts (part of which is lost in traversing the lead). Hence none of the two particles may have a kinetic energy above this limit. In some rare cases it has been possible to observe the creation of an electron pair (+ and −) within the gas of a Wilson cloud chamber, rather than in the lead. In that event the particles are not slowed down appreciably, and hence the sum of their kinetic energies must be equal to $h\nu - 2m_0c^2$, which value is generally well verified.

The inverse of the preceding phenomenon is the falling of an electron from a state of positive energy into one of the few unoccu-

---

[30] For a more complete bibliography and for greater details concerning this question and the annihilation of electrons, see Nos. 26 and No. 31 of the Bibliography, and the monograph by I. Curie and F. Joliot entitled *L'électron positif*. Paris: Hermann, 1934.

pied negative energy states, that is, the recombination of a negative electron with a positron, which both disappear while liberating a quantity of energy equal to $2m_0c^2$ in the form of radiation (plus their kinetic energy, which is generally negligible). This phenomenon, called *annihilation*, may take place essentially in two ways: (a) a positron encounters a free electron and combines with it, emitting two photons in opposite directions (not a single one, because the two electrons could not give it the required momentum). Therefore each of the two photons has an energy m_0c^2, or about half a million volts (where the kinetic energy of the two electrons is neglected); (b) a positron combines with an electron which is strongly bound to a nucleus. In case (b), the emission of a single photon is possible, of energy about equal to $2m_0c^2$, that is, of a frequency double that of case (a), and the recoil momentum is imparted to the nucleus. From a calculation of Fermi and Uhlenbeck, the probability of process (b) proves to be considerably smaller than that of process (a). Other theoretical possibilities for annihilation exist, but we confine ourselves here to considering only those two, of which the first is confirmed by notable experiments of various investigators.[31] Among them we cite only those of Thibaud and of Joliot.

These workers have observed independently that if positrons are incident upon lead, aluminum, or platinum, a radiation emerges from the bombarded metal whose absorption coefficient in lead corresponds to gamma rays of about half a million volts (more precisely, the limits assigned by Joliot are from 425,000 to 645,000 volts). Furthermore, a (necessarily inexact) measurement of the intensity of this radiation compared to the number of incident positrons has permitted Joliot to conclude that from 1.6 to 3 photons are emitted for every incident positron, with a good probability for the value 2 predicted from theory.

The phenomenon of annihilation provides a reason for the fact that positive electrons are so much more rarely observed than negative ones, and that they do not intervene in the conduction of electricity, in the thermionic effect, and in similar phenomena. In

---

[31] Chao, *Proc. Nat. Acad.* **16**, 431 (1930); *Phys. Rev.* **36**, 1519 (1930); *Proc. Roy. Soc.* **A135**, 206 (1931); Tarrant, *ibid.* **A136**, 662 (1932); Gray and Tarrant, *ibid.* **143**, 681 (1934); J. Thibaud, *Comptes Rendus* **197**, 1629 (1933); F. Joliot, *ibid.*, **197**, 1622 (1933), and *J. de Phys.* **V**, 299 (1934).

fact, a positron in matter (even in a rarefied gas) finds itself in the presence of a large number of negative electrons and has a strong tendency to combine with one of these. The probability for combination has been calculated by Dirac, who found that it increases, tending toward a maximum value upon a decrease in velocity of the positron, so that in practice the positron may survive only if it possesses a considerable velocity. The mean life of a slow positron in water, for example, proves to be 3.5×10^{-10} sec.

CHAPTER 15

Systems with Identical Particles

61. Identity of elementary particles. The intuitive concept of particles such as electrons, protons, and photons as material points is incorrect, not only because it attributes to them kinematic properties which do not correspond to reality (as was amply illustrated in Part I) but also because it assigns them an "individuality" which they do not have in nature. In order to understand this assertion better, let us think of a system containing several electrons (for example, an atom). If we let ourselves be guided by the corpuscular model and if we think of the electrons as tiny balls, we can always attach a meaning to the operation of interchanging two of these balls, even if they are identical (in the ordinary sense of the term). But if we analyze the meaning of this exchange operationally, we see that it is based on the possibility of attributing an individuality to the particle by "marking" it in some way (even so slightly as to leave its fundamental properties unchanged), or by following it uninterruptedly throughout the operation of exchange. But for electrons (and for any other elementary particle) it is *conceptually impossible* either to mark them or to follow them with continuity. For this reason too, then, the corpuscular model is inadequate for their representation. The property of "individuality" in this case does not correspond to any physical reality. The only reason why we are accustomed to attribute this property to elementary particles is that in our mind and our everyday language, it is indissolubly associated (just like the concepts of position, velocity, trajectory, and so on) with the word "particle," by which we usually describe electrons, protons, and the like. Now we have seen that as far as the kinematic concepts are concerned, the inadequacy of corpuscular terminology may be compensated for by the introduction of the uncertainty principle. Similarly, as far as the question of individuality is concerned, it is still possible to retain corpuscular terminology, but with the express convention that two elementary particles of the same kind are to be considered *identical* in a stricter sense of the word than is ordinarily implied. In this sense the expression "interchange of any two particles" loses its meaning, so that the characteristics of individuality are removed from the particle concept.[1] It is in this special sense that the *identity* of particles, with which we shall be dealing in this chapter, is to be understood. As we shall see, this property of elementary particles, which has no analogue among ordinary bodies, leads to singular consequences, which may not be intuitively interpreted by means of the corpuscular model but which in many cases are susceptible of experimental verification.

First let us consider systems with only two identical particles (such as, for instance, the helium atom). We shall then briefly extend the arguments to systems with any number of identical particles.

62. States of a system with two identical particles. Symmetry and antisymmetry. Let us consider a system composed of two identical particles (located in a given field), and let us indicate for short by q^1, p^1 the aggregate of coordinates and momenta of one of them, and by q^2, p^2 the corresponding quantities of the other. We also include the spin variable in the q, which we shall denote by σ and which might, for example, be the σ_z of the preceding chapter (hence, for example, q^1 stands for the group x_1, y_1, z_1, σ_1). Any observable of the system must have an expression $F(q^1, p^1, q^2, p^2)$ which is *symmetric* with respect to the two groups of variables; that is, the expression must remain the same when the q^1, p^1 are changed into the corresponding q^2, p^2. Otherwise, the interchange of the two particles would have physical significance. We shall indicate the symmetry of the function F by writing, for short,

$$F(1, 2) = F(2, 1),$$

---

[1] In order to clarify this statement by another example of entities devoid of individuality in the sense specified above, let us think of a vibrating string in which two systems of standing waves are simultaneously excited; one with frequency ν_1, amplitude A_1, phase φ_1, the other with frequency ν_2, amplitude A_2, phase φ_2. The phrase "interchange of the system of waves (ν_1, A_1, φ_1) with the system (ν_2, A_2, φ_2)" is completely devoid of meaning. In the same way, if there are two electrons in an atom, one in a quantum state defined by certain quantum numbers n_1, l_1, m_1, the other in a state defined by n_2, l_2, m_2, there is no meaning whatsoever in saying that the first electron is placed into the second state and the second electron into the first.

instead of

$$F(q^1, p^1, q^2, p^2) = F(q^2, p^2, q^1, p^1).$$

Hence the operator corresponding to a general observable F will also be symmetric. In particular, the expression for the energy will be symmetric, and hence also the Hamiltonian operator $\mathfrak{H}$. It follows that if $\psi_n(1, 2)$ is an eigenfunction belonging to the eigenvalue E_n, that is, if it satisfies the equation

$$\mathfrak{H}\psi_n(1, 2) = E_n\psi_n(1, 2), \tag{360}$$

then $\psi_n(2, 1)$ is also an eigenfunction belonging to the same eigenvalue, since the equation is still satisfied if the q^1 and the q^2 are interchanged in ψ_n.

Some important consequences are derived from this reasoning. First let us suppose that E_n is a simple eigenvalue. In that case $\psi_n(2, 1)$ cannot be essentially different from $\psi_n(1, 2)$; that is, we must have

$$\psi_n(2, 1) = c\psi_n(1, 2),$$

where c is a constant. If we then interchange q^1 and q^2 in this relation and multiply one equation by the other, we find $c^2 = 1$, or $c = \pm 1$. Hence an eigenfunction belonging to a simple eigenvalue either has the property

$$\psi_n(2, 1) = \psi_n(1, 2);$$

that is, it is *symmetric*, or else

$$\psi_n(2, 1) = -\psi_n(1, 2),$$

in which case it is said to be *antisymmetric*.

Let us now examine the case of degeneracy, in which we suppose that E_n is a multiple eigenvalue of order p, and let

$$\psi_n^{(1)}(1, 2), \psi_n^{(2)}(1, 2), \ . \ . \ . \ , \psi_n^{(p)}(1, 2) \tag{361}$$

be a fundamental set of mutually orthogonal eigenfunctions belonging to E_n (see §6 of Part II). It is known that this set may be replaced by any other system obtained from the previous one by an orthogonal transformation. We shall now show that we may choose the transformation in such a way that the system will be made up exclusively of symmetric and antisymmetric functions (not merely mutually orthogonal).

Let us consider, for instance, the first of the eigenfunctions (361); upon interchanging the q^1 and q^2, we obtain a new eigenfunction $\psi_n^{(1)}(2, 1)$ belonging to the same eigenvalue. The latter may coincide (to within a constant factor) with the same $\psi_n^{(1)}(1, 2)$; and then it may be shown, as we have just seen, that it is already symmetric or antisymmetric. Or it may coincide with another one of (361), for example $\psi_n^{(r)}(1, 2)$, and in that case we shall replace the pair of functions $\psi_n^{(1)}(1, 2)$, $\psi_n^{(r)}(1, 2)$ by the two (evidently nonzero) functions

$$\tfrac{1}{2}[\psi_n^{(1)}(1, 2) \pm \psi_n^{(1)}(2, 1)], \tag{362}$$

which are also linearly independent of each other and of the others; furthermore, one of them is symmetric and the other is antisymmetric. Finally, $\psi_n^{(1)}(2, 1)$ may not coincide with any of the functions (361), and then we may substitute for $\psi_n^{(1)}(1, 2)$ any one of the two functions (362), which are both independent of the others. Applying this procedure to all the eigenfunctions (361) in succession, we shall be able to replace them by an equal number of independent eigenfunctions, some of which are symmetric and others antisymmetric. Their orthogonality remains to be established. Let us consider the group of symmetric eigenfunctions. It is always possible to perform a linear transformation upon them such as to replace them by an equal number of mutually orthogonal independent eigenfunctions (see, for example, §6 of Part II) which will evidently also be symmetric. Similarly, the antisymmetric eigenfunctions may be replaced by an equal number of their linear combinations which are independent and mutually orthogonal (and evidently antisymmetric). Finally, we note that any symmetric function ψ_s and any antisymmetric function ψ_a are necessarily orthogonal, since the integral

$$\psi_s \cdot \psi_a = \int \psi_s^*(1, 2)\psi_a(1, 2)\, dq^1\, dq^2$$

must not change when the variables q^1 and q^2 are interchanged. On the other hand, this interchange only changes the sign of the integrand; hence the integral vanishes. Therefore we recognize that the p eigenfunctions, which are partly symmetric and partly antisymmetric, and which we have substituted for the function (361), are all mutually orthogonal.

If for any possible multiple eigenvalue we select the fundamental set of eigenfunctions in the manner described, we obtain a *complete* set of orthogonal eigenfunctions, which contains only eigenfunctions which are either symmetric or antisymmetric.

Let us now proceed to consider in general the possible (also nonstationary) states of the system. Let us assume, as is physically plausible, not only that the probability density $P = \psi\psi^*$ must be symmetric with respect to an exchange of the two particles, but also that if $\psi(1, 2)$ represents a *possible* state of the system, then $\psi(2, 1)$ will also represent a possible state. It may then be shown that $\psi(1, 2)$ must necessarily be either symmetric or antisymmetric. A ψ which does not have this property (such a ψ might be obtained by a linear combination of symmetric and antisymmetric eigenfunctions) does not represent a physically possible state of the system, though it may satisfy the time-dependent Schrödinger (or Dirac) equation.

Let $\psi(1, 2)$ be neither symmetric nor antisymmetric. If it represented a possible state, so would $\psi(2, 1)$, and hence their (nonzero) symmetric and antisymmetric combinations: $\psi_s = \frac{1}{2}[\psi(1, 2) + \psi(2, 1)]$, $\psi_a = \frac{1}{2}[\psi(1, 2) - \psi(2, 1)]$. But then any other combination of the type $\varphi = \psi_s + e^{i\delta}\psi_a$ (where δ is an arbitrary constant) would also represent a possible state. This condition is in contradiction with our first hypothesis, since the probability density of that state,

$$\rho = \varphi\varphi^* = \psi_s\psi_s^* + e^{i\delta}\psi_a\psi_s^* + e^{-i\delta}\psi_a^*\psi_s + \psi_a\psi_a^*,$$

is not symmetric in general. In fact, if the variables q^1 and q^2 are exchanged, ρ is changed into

$$\rho' = \psi_s\psi_s^* - e^{i\delta}\psi_a\psi_s^* - e^{-i\delta}\psi_a^*\psi_s + \psi_a\psi_a^*.$$

The difference $\rho - \rho' = 2(e^{i\delta}\psi_a\psi_s^* + e^{-i\delta}\psi_a^*\psi_s)$ does not vanish in general, in view of the arbitrary nature of δ. Hence if the state $\psi(1, 2)$ were possible, states would also be possible whose ρ is not symmetric; this condition is to be excluded.

Hence we have shown that the a priori possible states for the system fall into two classes which we shall call *symmetric states* and *antisymmetric states,* according to whether ψ is symmetric or antisymmetric (implying: with respect to an exchange of x_1, y_1, z_1, σ_1 and x_2, y_2, z_2, σ_2).

We shall now show that the system *can never pass from a sym-*

metric state to an antisymmetric state, or vice versa; and that consequently, in the evolution of a given system, states of only one of these two classes may occur (as we shall see later, for electrons and protons the antisymmetric states are the only ones which arise). In fact, from the fundamental equation

$$\mathfrak{H}\psi = -\frac{h}{2\pi i}\frac{d\psi}{dt}, \tag{363}$$

we find that the increment of ψ in the time dt is

$$d\psi = -\frac{2\pi i}{h}\mathfrak{H}\psi\,dt. \tag{364}$$

Now since $\mathfrak{H}$ is a symmetric operator, if ψ is symmetric (or antisymmetric) at a given instant t, the same will be true for $d\psi$, and hence also for $\psi + d\psi$. Therefore ψ at the time $t + dt$ has the same symmetry character which ψ had at time t.

63. Note on the extension to N identical particles. If the system consists of N identical particles, and if we denote by q^i, p^i the coordinates and momenta of the ith particle (including the spin variables), it is obvious that any quantity having physical significance, and in particular the probability density $|\psi|^2$, must be symmetric with respect to all the particles; that is, it must remain unchanged when, in any manner whatsoever, the indices characterizing the variables are interchanged. In particular, from the fact that the Hamiltonian $\mathcal{H}$ possesses this property, we deduce, by the same procedure which was used in the case of two particles, that if in an eigenfunction $\psi_n(q^1, q^2, \ldots q^N)$ we permute the indices of the variables in any manner, then we obtain an eigenfunction belonging to the same eigenvalue. If the eigenvalue is simple, this eigenfunction must coincide with the original one to within a constant factor, which (as may be shown as in the previous case) can be only ± 1. We shall prove that this factor is either $+1$ for all permutations (*symmetric* function), or $+1$ for the even permutations[2] and -1 for the odd permutations (*antisymmetric* function).

First let us consider a *transposition* (r, s), that is, the exchange of only two indices, r and s, and let us indicate by C_{rs} the factor ($= \pm 1$)

---

[2] A permutation is even or odd according to whether it may be obtained by an even or odd number of transpositions, that is, of exchanges of only two elements at a time.

by which the function is multiplied as a consequence of this exchange. We then note that the transposition (r, s) may also be obtained by the successive application of the transpositions (r, t), (t, s), (r, t), where t is any other index. Therefore between the corresponding factors C we shall have the relation $C_{rs} = C_{rt}C_{ts}C_{rt} = C_{ts}C_{rt}^2 = C_{ts}$. But if $C_{rs} = C_{ts}$ no matter what the indices r, s, t, then all the C are equal to one another. Hence they are either all equal to $+1$, in which case any transposition, and therefore also any permutation, leaves the function unaltered; or else they are all equal to -1, and then an even number of transpositions will leave the function unaltered, and an odd number will change its sign—that is, we are dealing with an antisymmetric function. Note that in this argument we have not made use of the equation which ψ_n satisfies; and hence it holds not only for an eigenfunction but also for any function of the coordinates which, upon any transposition of the indices of the latter, is multiplied by ± 1.

If the eigenvalue is multiple, a permutation of the indices in ψ_n will change it into a different eigenfunction, in general independent of the former one. However, it may be shown that for any multiple eigenvalue of order p there exists a set of p (independent and orthogonal) eigenfunctions of which one is symmetric, one antisymmetric, and the others more complicated in behavior. However, these do not interest us here because, as we shall see, they do not represent physically possible states.[3]

Let us now consider some general state, even nonstationary, represented by a certain ψ. We assume (as in the case of two particles) that when the coordinates of any two particles are interchanged, the ψ is changed into another function representing a physically possible state. Hence we may apply the reasoning of the preceding section to each of these exchanges or transpositions, and we reach the conclusion that ψ must be either symmetric or antisymmetric with respect to any transposition. Therefore (since we have just seen that a function may not be symmetric for some transpositions and antisymmetric for others) *ψ must be either symmetric or antisymmetric (with respect to any permutation of the indices)*. Correspondingly, we shall speak of *symmetric states* and *antisymmetric states;*

---

[3] The general study of the symmetry properties of the eigenfunctions of N particles was made with the methods of group theory by E. Wigner, *Zeits. f. Physik* **40**, 883 (1927).

states which do not fall into one of these two categories are not possible.

The arguments presented at top of page 451 may be extended immediately to the case of several particles, in order to show that the system cannot pass in any way from a symmetric to an antisymmetric state, or vice versa.

64. The Pauli principle and the antisymmetry principle. In the Bohr-Sommerfeld theory of the atom, as we know, three quantum numbers were assigned to each electron—the *principal quantum number* n_i, *the azimuthal quantum number* l_i, and the *magnetic quantum number* m_i—which completely define its orbit. If we add to these, in accordance with the hypothesis of the spinning electron, the *spin* quantum number s_i (capable of having only two values, corresponding to the two opposite orientations which the spin may take), the quantum state of each of the electrons in the atom will be specified by a *quadruplet* of quantum numbers. These numbers are involved in all laws which determine the behavior of the electron in question (for example, in the selection rules) and are therefore closely related to the chemical and spectroscopic phenomena of the outermost electrons (valence and emission electrons) and to the emission and absorption of X rays for the inner electrons. Hence it is possible to check their values experimentally.

Actually, in §56 of Part II, we have defined the quantum numbers only for a single electron in a central field. In an atom with several electrons, their interaction complicates the problem. In first approximation we can, as was seen in §59 of Part II, idealize the influence exerted upon the ith electron by the remaining ones by considering that a central force is added to the force due to the nucleus, whereby we are led back to the previous situation. But even independently of any approximation, it is always possible to define a group of four quantum numbers n_i, l_i, m_i, s_i for each electron in the following manner. Let us suppose that the atom is placed in a magnetic field whose value, starting from zero, increases slowly until the forces which it exerts upon the electrons become predominant over all their mutual interactions, and over the action of the orbital magnetic field upon the spins. It may be shown that if the atom was originally in a given quantum state, it will remain there even in a strong magnetic field. But in that field each electron may be quantized separately, and hence its four quantum numbers may be assigned in the usual way. Since the system was brought continuously from a zero magnetic field to a very strong field, whereas the quantum numbers cannot vary with continuity, we must attribute to each electron in the atom not subject to the field, the same quantum numbers which belong to it in a sufficiently intense magnetic field.

Now it was discovered empirically (for the first time by Pauli while he was studying the spectra of the alkali metals)[4] that *in an atom there never are two electrons having the same group of four quantum numbers.*[5] This fact cannot be deduced from any other law of quantum theory; but since it is consistently borne out by all atoms (and also by more complex systems, such as molecules), it was assumed as one of the fundamental principles of atomic physics under the name of *Pauli principle* or *exclusion principle,* and has proved extremely fertile in consequences in full accord with experience.[6] It may be said that the entire theoretical interpretation of spectra is based upon this principle and constitutes an impressive confirmation thereof. In the present volume, we shall have occasion to cite only one of these applications (see §67).

The Pauli principle has furnished the key for the interpretation of one of the most fundamental laws of nature, namely, of the periodic system of Mendelyeev. For a detailed presentation of this interpretation we refer to a work on spectroscopy (see footnote, page 252). We confine ourselves here to a brief schematic and simplified outline. If the Pauli principle did not hold, each electron would tend to be located in the orbit of least energy, which is (at least for the lighter elements, to which we restrict ourselves here) the orbit with $n = 1$, $l = 0$, $m = 0$. Therefore all the electrons of the atom in the ground state would have these quantum

---

[4] *Zeits. f. Physik* **31,** 765 (1925).

[5] The fourth quantum number s can take on only two values. Thus when the values of the three quantum numbers n, l, m (called "orbital" quantum numbers) are fixed, there may be at most *two* electrons having these three orbital quantum numbers (or, as is said for short, having this "orbit"). This is the form in which Pauli stated his principle for the first time. The two electrons having the same orbit must then differ in their quantum number s, that is, they must have their spins in opposite directions.

[6] The Pauli principle is assumed for any system containing several electrons; for example, "Fermi statistics," valid for an electron gas, and in particular for the conduction electrons in metals, is based upon this principle. There is reason to believe that the Pauli principle holds also for protons and some nuclei. For others, however (for instance, for α-particles) it is not valid.

numbers (or, in spectroscopic terminology, they would be in the 1*s*-state), and the optical and chemical properties would vary progressively from one atom to the next, corresponding to the increase of the number of these electrons. In particular, the fundamental term of the spectrum would always be 1*s*. All this is in absolute contrast with reality. Instead, according to the Pauli principle, in the afore-mentioned orbit there cannot be more than two electrons. Therefore in the lithium atom (atomic number $Z = 3$, hence three electrons), two electrons find themselves in the first orbit, but the third must be located in the next orbit (in order of increasing energy), which is $n = 2$, $l = 0$, $m = 0$. This electron is therefore less tightly bound to the nucleus than the others, and hence acts as valence electron (hence Li is monovalent) and as emission electron (hence the fundamental term is 2*s*). In Be ($Z = 4$) a second electron occupies the second orbit (hence valence of two, and fundamental term 2*s*). In B ($Z = 5$), the first two orbits being completely occupied, the fifth electron will go into an orbit $n = 2$, $l = 1$, which explains why the ground term of boron is 2*p*. Thus, by adding one electron at a time to the lowest energy state which remains free, compatible with the Pauli principle (and of course, increasing the charge of the nucleus by one unit every time), we determine the quantum state of the various electrons in successive elements. In particular, the quantum state of the electron added last defines the ground (fundamental) term of the spectrum. The predictions obtained in this way are in remarkable accord with the experimentally observed properties of the different elements. The existence of the so-called "periods" is explained by the fact that the successively added electrons occupy (as a rule) first the orbits with $l = 0$, then with $l = 1$, then with $l = 2, \ldots,$ up to $n - 1$, which is the maximum l compatible with a given n. After this we pass to the orbits with $n + 1$, starting once again with $l = 0$, $l = 1$, and so on. And since it follows from a simple calculation that at a fixed n there are n^2 orbits, that is, $2n^2$ places, we shall have a first period of $2 \cdot 1^2 = 2$ elements (H and He), a second one with $2 \cdot 2^2 = 8$, and so forth.

For further particulars, see, for example, No. 27 of the Bibliography.

In the statement of the Pauli principle, we have thus far used the terminology of the Bohr-Sommerfeld theory. But in wave mechanics, this principle has received an even simpler and more general formulation by the work of Dirac, since it also applies to nonstationary states. This formulation states: *in a system containing several electrons, only the antisymmetric states are possible* (in the sense explained on page 453). But neither does wave mechanics provide any proof of this principle (which we shall call *antisymmetry principle*), and hence it must be considered a postulate. However, wave mechanics shows (as we have seen) that if a system obeys this principle at a certain instant, it will satisfy it forever.

We shall now show that the Pauli exclusion principle follows as a necessary consequence from Dirac's antisymmetry principle. Let us suppose that the system is in a stationary state, and let us place it in a magnetic field whose intensity, which is very weak at first, we shall increase up to a value such that we may neglect the interactions between the various electrons, as we have pointed out on page 453. Let $\psi_{n_i}(q^i)$ be the eigenfunction of the ith electron (n_i stands for the group of four numbers n_i, l_i, m_i, s_i, and q^i stands for x_i, y_i, z_i, σ_i). Since the different electrons are dynamically independent, the equation of the total system has for eigenfunction the product of the eigenfunctions corresponding to the separate electrons which compose the system (see §20), that is, $\psi_{n_1}(q^1)$, $\psi_{n_2}(q^2) \dots \psi_{n_N}(q^N)$. But since the equation is symmetric with respect to the N groups of four coordinates, it will also be satisfied by any function (corresponding to the same eigenvalue) which is obtained from the previous one by any permutation of the q^i, namely by

$$\psi_{n_1}(q^{r_1})\psi_{n_2}(q^{r_2}) \dots \psi_{n_N}(q^{r_N}), \tag{365}$$

where $r_1, r_2, \dots r_N$ represents any permutation of the numbers $1, 2, \dots N$. The general solution will be a linear combination of all the solutions obtained in this manner. Of these combinations, one will be symmetric, which is the sum of all the $N!$ products of the type (365):

$$\psi_s = \Sigma\psi_{n_1}(q^{r_1})\psi_{n_2}(q^{r_2}) \dots \psi_{n_N}(q^{r_N}) \tag{366}$$

(the summation being extended over all permutations $r_1, r_2, \dots r_N$); and one combination will be antisymmetric and may be written in the form of a determinant as follows:

$$\psi_a = \begin{vmatrix} \psi_{n_1}(q^1) & \psi_{n_1}(q^2) & \cdots & \psi_{n_1}(q^N) \\ \psi_{n_2}(q^1) & \psi_{n_2}(q^2) & \cdots & \psi_{n_2}(q^N) \\ \cdots & \cdots & \cdots & \cdots \\ \psi_{n_N}(q^1) & \psi_{n_N}(q^2) & \cdots & \psi_{n_N}(q^N) \end{vmatrix}. \tag{367}$$

[In fact, the development of this determinant is just the sum of all expressions of the type (365), preceded by a plus or minus sign, according to whether the permutation $r_1, r_2, \ldots r_N$ is even or odd. The expression is then evidently antisymmetric, since the exchange of any two q, which is equivalent to an exchange of two columns of the determinant, only changes the sign of the latter.] If we now assume the antisymmetry principle, we must adopt ψ_a as the only eigenfunction possible among all those mentioned above. Under this assumption, if there are two or more electrons in the system with the same quantum numbers, there would be two or more equal rows in the determinant, and hence ψ_a would be identically zero. Hence the antisymmetry principle leads to the exclusion of the possibility of several electrons in the same quantum state, and contains the Pauli principle.

Conversely, the antisymmetry principle is a necessary consequence of the Pauli principle, provided the latter is understood in a somewhat more restricted sense than that presented previously. This follows if we assume, not only that two electrons cannot exist in the same quantum state of the system, but also that it is not possible to bring the system into such a state, no matter what the forces to which it is subjected (provided, of course, that they act upon all electrons in a symmetric manner). In fact, if ψ were symmetric instead of antisymmetric (we have already seen in §63 that no other cases can present themselves) it would be possible, by an appropriate (symmetric) Hamiltonian, to make it evolve in such a way as to coincide, after a certain time, with the symmetric eigenfunction ψ_s given by (366), in which the indices $n_1, n_2, \ldots n_N$ have been fixed in any manner whatever, even with any number of them equal to each other.

Therefore not only does the antisymmetry principle contain the Pauli principle; it also constitutes the necessary condition to ensure its permanent validity in a system, no matter in what physical circumstances the system finds itself. In that sense, the antisymmetry principle may be considered equivalent to the Pauli principle and is actually often called by that name.

65. Approximate calculation of the states of a system with two identical particles. In a system consisting of two particles, each of the particles may be considered to be acted upon by: (a) forces independent of the presence of the second particle; (b) (electric and magnetic) forces due to the other particle. In many cases, these latter forces may be considered small with respect to the former, so that we may neglect them in a preliminary approximation (zero-order approximation); we introduce them later into a first approximation by the perturbation method and, if need be, into higher approximations. As we shall see, we may treat the helium atom and the hydrogen molecule by this method, these two cases being the most interesting applications of this theory.[7]

First let us deal with the zero-order approximation; that is, let us neglect the interaction between the two particles. We can then consider each of them as belonging to a different system (ignoring the spatial superposition of the two systems); and, in a stationary state, we may attribute to the first particle an eigenfunction $\psi_{n_1}(1)$, to the second an eigenfunction $\psi_{n_2}(2)$ (where, as usual, n_1, n_2 represent two groups of four quantum numbers, and 1, 2 represent two groups of four coordinates). These eigenfunctions satisfy the two equations

$$\mathfrak{H}_1^0\psi_{n_1} = E_{n_1}\psi_{n_1}, \qquad \mathfrak{H}_2^0\psi_{n_2} = E_{n_2}\psi_{n_2}, \tag{368}$$

where we have denoted by $\mathfrak{H}_1^0$ the Hamiltonian relative to the first particle, and by $\mathfrak{H}_2^0$ that of the second particle (neglecting their interaction). These operators, of course, have the same form, and the second differs from the first only by the mathematical substitution of the letters x_2, y_2, z_2, σ_2 for x_1, y_1, z_1, σ_1. Hence $\psi_{n_1 s_1}$ and $\psi_{n_2 s_2}$ are obtained, in substance, by solving one and the same eigenvalue problem. It suffices to write the following equation, which

---

[7] In order to obtain greater precision in the zero approximation, we may also include part of the interaction force, representing it schematically as a central field acting on each particle and equal, approximately, to the average value of the force exerted upon it by the other particle (*screening action;* see §59 of Part II). The perturbation then reduces to the difference between the true interaction and this force. This result does not substantially alter anything of what follows.

is the same as (368) but lacks indices,

$$\mathfrak{H}^0\psi(x, \sigma) = E\psi(x, \sigma), \tag{369}$$

and then to take one of its particular eigenfunctions ψ_{n_1} and to affix the index 1 to the variables in it. Similarly, we take another particular eigenfunction ψ_{n_2} (which may possibly coincide with the former) and write the variables with index 2. E_{n_1}, E_{n_2} are the corresponding (possibly equal) eigenvalues. If we now consider the two particles as belonging to a single system (but still neglect mutual interactions), we know from §20 that to the system there belongs the operator $\mathfrak{H}_1^0 + \mathfrak{H}_2^0$; and that its state, when the first particle is in the state n_1 and the second particle is in the state n_2, is expressed by the product of the two respective eigenfunctions, that is, by the following ψ:

$$\psi_1^0 = \psi_{n_1}(q^1)\psi_{n_2}(q^2). \tag{370}$$

The corresponding energy is

$$E^0 = E_{n_1} + E_{n_2}. \tag{371}$$

Now, because of the fundamental facts established in §62, if we exchange the variables q^1 and q^2 in function (370), we again obtain an eigenfunction of the system, belonging to the same eigenvalue. Indicating it by ψ_2^0, we have

$$\psi_2^0 = \psi_{n_1}(q^2)\psi_{n_2}(q^1). \tag{372}$$

Let us now suppose that the group of four indices n_1 is not identical to the group n_2, that is, that the two particles are in different states. Then ψ_2^0 will result different from ψ_1^0, and hence the energy level (371) will be double (independently of other possible degeneracies, which we are ignoring). Therefore: *in the zero-order approximation there occurs a special degeneracy, deriving from the identity of the two particles, by virtue of which all energy levels turn out to be double, except for those corresponding to quantum states which are equal for both particles.* This degeneracy is called *exchange degeneracy* or *resonance degeneracy,* for reasons which will be explained shortly.

Let us now pass to the first approximation, taking into account the interaction between the two particles, which we shall introduce as a perturbation, following the methods of Chapter 13. The

Hamiltonian of the system will then be written in the form

$$\mathfrak{H}_1^0 + \mathfrak{H}_2^0 + \mathfrak{L}, \tag{373}$$

where the operator $\mathfrak{L}$ corresponds to the interaction energy of the two particles, and hence contains the variables q^1 and q^2 symmetrically; for example, if the interaction reduces to the electrostatic repulsion, then $\mathfrak{L} = e^2/r_{12}$, where

$$r_{12} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}.$$

However, if we take the magnetic interactions into account, $\mathfrak{L}$ will also contain the spin variables σ_1 and σ_2, but always in a symmetric manner. Because of the exchange degeneracy, we must apply the formulas found in §39, and hence we must first construct the "perturbation matrix," whose elements are

$$\left.\begin{aligned} L_{11} &= \sum_{s_1s_2} \int \psi_1^{0*}\mathfrak{L}\psi_1^0\,dS, & L_{12} &= \sum_{s_1s_2} \int \psi_1^{0*}\mathfrak{L}\psi_2^0\,dS, \\ L_{21} &= \sum_{s_1s_2} \int \psi_2^{0*}\mathfrak{L}\psi_1^0\,dS, & L_{22} &= \sum_{s_1s_2} \int \psi_2^{0*}\mathfrak{L}\psi_2^0\,dS, \end{aligned}\right\} \tag{374}$$

where every integral is sixfold and dS stands for $dx_1\,dy_1\,dz_1\,dx_2\,dy_2\,dz_2$, whereas the summations are double and stand for the integration with respect to the two discontinuous variables s_1, s_2 (spin quantum numbers), each of which takes on only two values. Then we note that it is legitimate to interchange, in each of these expressions, the designation of the variables of integration and of the indices of summation, without changing the result. In fact, if we interchange the variables q^1 and q^2, ψ_1^0 changes into ψ_2^0, and vice versa; and hence L_{11} changes into L_{22}, L_{12} into L_{21}, and vice versa. Therefore

$$L_{11} = L_{22}, \qquad L_{12} = L_{21}. \tag{375}$$

In first approximation, the perturbation ϵ' of the value of the energy is found, as we have seen in §39, by solution of the equation

$$\begin{vmatrix} L_{11} - \epsilon' & L_{12} \\ L_{21} & L_{22} - \epsilon' \end{vmatrix} = 0, \tag{376}$$

which, by virtue of (375), yields the following two values for ϵ' (the reason for the notation ϵ_s', ϵ_a' will appear presently):

$$\epsilon_s' = L_{11} + L_{12}, \qquad \epsilon_a' = L_{11} - L_{12}. \tag{377}$$

Hence the level (371) will split into the two levels

$$E_s = E_{n_1} + E_{n_2} + L_{11} + L_{12}, \tag{378}$$

$$E_a = E_{n_1} + E_{n_2} + L_{11} - L_{12}. \tag{378$'$}$$

Let us now seek the eigenfunctions of the zero-order approximation corresponding to these eigenvalues. They are given by (see §39)

$$\left.\begin{aligned} \psi_s^0 &= c_{11}^0\psi_1^0 + c_{12}^0\psi_2^0, \\ \psi_a^0 &= c_{21}^0\psi_1^0 + c_{22}^0\psi_2^0, \end{aligned}\right\} \tag{379}$$

where the coefficients c^0 are obtained from (185′), which in the present case are written, taking $k = 1$ (for $k = 2$ one would obtain an equivalent system),

$$\begin{aligned} c_{11}^0(L_{11} - \epsilon_s') + c_{12}^0 L_{12} &= 0 \\ c_{21}^0(L_{11} - \epsilon_a') + c_{22}^0 L_{12} &= 0, \end{aligned}$$

or, if we take (375) and (377) into account and assume that we have $L_{12} \neq 0$,[8]

$$c_{11}^0 = c_{12}^0, \qquad c_{21}^0 = -c_{22}^0.$$

Upon normalizing ψ_s and ψ_a, we find that the modulus of these coefficients must be $1/\sqrt{2}$, so that we may write

$$\psi_s^0 = \frac{1}{\sqrt{2}}(\psi_1^0 + \psi_2^0), \qquad \psi_a^0 = \frac{1}{\sqrt{2}}(\psi_1^0 - \psi_2^0) \tag{380}$$

or

$$\psi_s^0 = \frac{1}{\sqrt{2}}[\psi_{n_1}(q^1)\psi_{n_2}(q^2) + \psi_{n_1}(q^2)\psi_{n_2}(q^1)], \tag{381}$$

$$\psi_a^0 = \frac{1}{\sqrt{2}}[\psi_{n_1}(q^1)\psi_{n_2}(q^2) - \psi_{n_1}(q^2)\psi_{n_2}(q^1)]. \tag{381$'$}$$

As can be seen, the first of these eigenfunctions is symmetric; the second, antisymmetric. This result could have been foreseen because of what was said in §62. In the case where the particles are electrons or protons, the Pauli principle excludes the first of these, and only ψ_a^0 remains to represent a possible state of the system. Hence (378′) alone represents all physically possible energy levels.

---

[8] The case $L_{12} = 0$ is realized when the two probability clouds represented by the functions ψ_{n_1}, ψ_{n_2} have no points in common: in that case a separate region of space is assigned to each of the two electrons, and it is as if each of them had its own individuality. Hence no exchange phenomenon occurs.

The case, excluded until now, in which both particles have the same quantum numbers ($n_1 = n_2$), does not give rise to degeneracy. Hence there exists a single eigenfunction in that case, and it is symmetric. Evidently we must exclude these states when dealing with particles obeying the Pauli principle.

We shall now briefly outline an argument which, though of purely formal value, explains the name of *exchange* or *resonance* phenomena, which is generally given to phenomena characteristic of systems with equal particles.[9] Let us suppose that the two particles in question, though dynamically equal, may be distinguished in some manner by means of a "mark" which does not alter their mechanical properties (a supposition which is evidently fictitious and devoid of physical significance). Under this hypothesis it is possible to distinguish between the case in which particle 1 is in state n_1 and particle 2 in state n_2, and the case in which they are interchanged, that is, between the two states of the system represented by ψ_1^0 and ψ_2^0 in the zero-order approximation. Besides ψ_s and ψ_a, there will then also be admissible any of their linear combinations

$$\psi = c_s\psi_s + c_a\psi_a, \tag{382}$$

with c_s and c_a constant. Since to ψ_s there corresponds the eigenvalue E_s, and to ψ_a the eigenvalue E_a in first approximation, these eigenfunctions may be written in the form

$$\psi_s = u_s\, e^{-\frac{2\pi i}{h}E_st}, \qquad \psi_a = u_a\, e^{-\frac{2\pi i}{h}E_at},$$

where u_s and u_a are functions of the coordinates but not of t, and E_s and E_a are given by (378) and (378′). Substituting in (382), and putting

$$C_s = c_s\, e^{-\frac{2\pi i}{h}(L_{11}+L_{12})t}, \qquad C_a = c_a\, e^{-\frac{2\pi i}{h}(L_{11}-L_{12})t}, \tag{383}$$

we have

$$\psi = (C_su_s + C_au_a)\, e^{-\frac{2\pi i}{h}E^0t}. \tag{384}$$

It is apparent that C_s and C_a are not constants like c_s and c_a, but vary periodically and slowly. In fact, their frequency $(L_{11} \pm L_{12})/h$ is small compared with the frequency E^0/h of the principal factor of ψ, since the perturbation is small compared with the energy.

---

[9] See W. Heisenberg, *Zeits. f. Physik* **38**, 411 (1926).

If the solution (384) is considered for a time that is short compared with the period of variation of C_s and C_a, so as to enable us to regard them as constants, this solution can be approximately identified with the linear combination

$$\psi = C_s\psi_s^0 + C_a\psi_a^0 \tag{385}$$

of the solutions of the zero-order approximation (380). In other words, our ψ may be considered to be a linear combination, with slowly varying coefficients, of ψ_s^0 and ψ_a^0, and hence also of ψ_1^0 and ψ_2^0. In fact we have, after substituting (380) into (384),

$$\psi = \frac{1}{\sqrt{2}}(C_s + C_a)\psi_1^0 + \frac{1}{\sqrt{2}}(C_s - C_a)\psi_2^0.$$

Let us now suppose that we have found out, at time $t = 0$, that particle 1 is in state n_1 and particle 2 in state n_2, that is, that ψ is represented (approximately) by ψ_1^0. Then we get from the preceding equation and from (383), setting $t = 0$ and $\psi = \psi_1^0$, that $c_s = c_a = 1/\sqrt{2}$, and hence for any time t we may write

$$\psi = e^{-\frac{2\pi i}{h}L_{11}t}\left[\psi_1^0 \cos\frac{2\pi}{h}L_{12}t + i\psi_2^0 \sin\frac{2\pi}{h}L_{12}t\right]. \tag{386}$$

Now the square of the modulus of the coefficient of ψ_1^0, or $\cos^2 (2\pi/h)L_{12}t$, represents the probability of finding the system in the state ψ_1^0 at time t, that is, particle 1 in state n_1 and particle 2 in state n_2, while $\sin^2 (2\pi/h)L_{12}t$ analogously represents the probability of finding the two particles interchanged. We see that at time $\tau/2$, where $\tau = h/2L_{12}$, the two particles are found to be interchanged with certainty; at time τ they are again as they were at time $t = 0$; and so forth. At intermediate instants, the state of the system is such that we do not know which particle is in state n_1 and which in state n_2. The smaller the coupling between the two particles, as measured by the integral L_{12} (the "exchange integral"), the larger is the period τ with which the two probabilities oscillate.

This curious exchange phenomenon is analogous, as far as its analytic aspect is concerned, to the periodic energy exchange which takes place between two oscillators of equal frequency which are loosely "coupled"; for example, between two pendula of the same length hung from a common horizontal wire which is not too taut. This mechanical analogy has suggested the name "resonance."

66. Separation of the spin coordinates from the position coordinates. In many cases in a system with two electrons, the forces due to the spins are negligible in first approximation. Formally, this statement means that the spin variables appear neither in $\mathfrak{H}^0$ nor in the main part of the interaction term $\mathfrak{L}$, but only in a term of higher order, which is being neglected in a first approximation. When these conditions are realized, it is said that there is *Russell-Saunders coupling* between the two electrons.

The first of these hypotheses enables us to treat (as was seen in §45 in the Pauli approximation) the unperturbed problem by separating the spin variable σ from the position variables, that is, by writing a general solution of (369) in the form

$$\psi_{n_i s_i} = \Psi_{n_i}(x_i)\varphi_{s_i}(\sigma_i), \tag{387}$$

where $i = 1, 2$, and n_i now represents only the group of three *orbital* quantum numbers of the ith electron, and the spin quantum number s_i is considered separately. Similarly, x_i stands for the three position coordinates x_i, y_i, z_i. The factor Ψ_{n_i} satisfies the Schrödinger equation

$$\mathfrak{H}^0\Psi_{n_i} = E_{n_i}\Psi_{n_i}, \tag{388}$$

and the spin eigenfunction φ_{s_i} reduces essentially to a group of two constants α_{s_i} and β_{s_i} (corresponding to the two values ± 1 of the variable σ_i, respectively), so that we may write, designating by s an index which can take on the two values ± 1,

$$\varphi_s(\sigma) = \begin{vmatrix}\alpha_s\\ \beta_s\end{vmatrix}.$$

Note that since s can take only two values, there exist only two "functions φ," or pairs (α_s, β_s). Let us assume these "functions" to be orthonormal; that is,

$$\sum_\sigma \varphi_{s_1}(\sigma)\varphi^*_{s_2}(\sigma) = \alpha_{s_1}\alpha^*_{s_2} + \beta_{s_1}\beta^*_{s_2} = \delta_{s_1s_2}. \tag{389}$$

After these preliminaries, the two eigenfunctions of zero-order approximation, namely, the symmetric and antisymmetric eigenfunctions (381) and (381′), may now be written as follows, making use of (387):

$$\left.\begin{aligned}\psi_s^0 &= \frac{1}{\sqrt{2}}[\Psi_{n_1}(x_1)\Psi_{n_2}(x_2)\varphi_{s_1}(\sigma_1)\varphi_{s_2}(\sigma_2)\\&\qquad\qquad + \Psi_{n_1}(x_2)\Psi_{n_2}(x_1)\varphi_{s_1}(\sigma_2)\varphi_{s_2}(\sigma_1)],\\ \psi_a^0 &= \frac{1}{\sqrt{2}}[\Psi_{n_1}(x_1)\Psi_{n_2}(x_2)\varphi_{s_1}(\sigma_1)\varphi_{s_2}(\sigma_2)\\&\qquad\qquad - \Psi_{n_1}(x_2)\Psi_{n_2}(x_1)\varphi_{s_1}(\sigma_2)\varphi_{s_2}(\sigma_1)].\end{aligned}\right\}$$

With the eigenfunctions of position Ψ we now form the following combinations, of which the first is symmetric and the second antisymmetric:

$$\left.\begin{aligned}\mathcal{S}(x_1, x_2) &= \tfrac{1}{2}[\Psi_{n_1}(x_1)\Psi_{n_2}(x_2) + \Psi_{n_1}(x_2)\Psi_{n_2}(x_1)],\\ \mathcal{A}(x_1, x_2) &= \tfrac{1}{2}[\Psi_{n_1}(x_1)\Psi_{n_2}(x_2) - \Psi_{n_1}(x_2)\Psi_{n_2}(x_1)],\end{aligned}\right\} \tag{390}$$

and similarly with the spin eigenfunctions:

$$\left.\begin{aligned}s(\sigma_1, \sigma_2) &= \tfrac{1}{2}[\varphi_{s_1}(\sigma_1)\varphi_{s_2}(\sigma_2) + \varphi_{s_1}(\sigma_2)\varphi_{s_2}(\sigma_1)],\\ a(\sigma_1, \sigma_2) &= \tfrac{1}{2}[\varphi_{s_1}(\sigma_1)\varphi_{s_2}(\sigma_2) - \varphi_{s_1}(\sigma_2)\varphi_{s_2}(\sigma_1)].\end{aligned}\right\} \tag{391}$$

Then ψ_s^0 and ψ_a^0 may be expressed in terms of these new combinations, and they become

$$\psi_s^0 = \sqrt{2}\,(\mathcal{S}s + \mathcal{A}a), \tag{392}$$
$$\psi_a^0 = \sqrt{2}\,(\mathcal{S}a + \mathcal{A}s). \tag{392′}$$

Note now that since we neglect the forces due to the spins, the eigenvalues will result independent of the spin quantum numbers s_1, s_2, depending only on the remaining quantum numbers n_1, n_2. From this condition is derived a further degeneracy, which we shall now investigate.

If n_1 and n_2 are fixed, four possibilities exist for s_1 and s_2:

$$\left.\begin{aligned}&(1)\ s_1 = +1, && s_2 = +1\\ &(2)\ s_1 = -1, && s_2 = -1\\ &(3)\ s_1 = +1, && s_2 = -1\\ &(4)\ s_1 = -1, && s_2 = +1,\end{aligned}\right\} \tag{393}$$

to which there correspond an equal number of symmetric eigenfunctions of the type (392), which we shall denote by ψ_s^1, ψ_s^2, ψ_s^3, ψ_s^4, respectively, and an equal number of antisymmetric eigenfunctions of the type (392′), which we shall call ψ_a^1, ψ_a^2, ψ_a^3, ψ_a^4. They all correspond to the eigenvalue $E_{n_1} + E_{n_2}$ in zero-order approximation.

However, when the perturbation $\mathfrak{L}$ is introduced, these eigenfunctions must be replaced (see §39) by an equal number of appropriately selected linear combinations, which we designate by Ψ_s^i, Ψ_a^i $(i = 1, 2, 3, 4)$. The Ψ_s^i will be formed by the ψ_s alone and the Ψ_a^i by the ψ_a alone, since, as we have said in §62, a linear combination of a ψ_s with a ψ_a is physically not permissible. However, since we are dealing with electrons, which obey the Pauli principle, we must exclude the symmetric eigenfunctions Ψ_s^i, and hence we shall deal only with the Ψ_a^i from now on. They are given by the formulas

$$\Psi_a^i = \sum_j c^{ij}\psi_a^j,$$

where the coefficients c^{ij} are obtained (see §39) by means of the four systems of linear equations

$$\left.\begin{aligned}(L^{11} - \epsilon^i)c^{i1} + L^{12}c^{i2} + L^{13}c^{i3} + L^{14}c^{i4} &= 0,\\ L^{21}c^{i1} + (L^{22} - \epsilon^i)c^{i2} + L^{23}c^{i3} + L^{24}c^{i4} &= 0,\\ L^{31}c^{i1} + L^{32}c^{i2} + (L^{33} - \epsilon^i)c^{i3} + L^{34}c^{i4} &= 0,\\ L^{41}c^{i1} + L^{42}c^{i2} + L^{43}c^{i3} + (L^{44} - \epsilon^i)c^{i4} &= 0,\end{aligned}\right\} \qquad (394)$$

and the elements L^{jl} of the perturbation matrix are given by

$$L^{jl} = \sum_{\sigma_1\sigma_2} \int \psi_a^{j*}\mathfrak{L}\psi_a^l\, dS. \qquad (395)$$

ϵ_i is one of the roots obtained by equating the determinant of the coefficients to zero, and the eigenfunction Ψ_a^i belongs to the eigenvalue

$$E_{n_1} + E_{n_2} + \epsilon^i.$$

In order to calculate L^{jl}, let us use (391) to find the spin eigenfunctions which correspond to the four pairs of values (393) for s_1 and s_2. We obtain

$$\left.\begin{aligned}&s_1 = \varphi_1(\sigma_1)\varphi_1(\sigma_2) \qquad\qquad s_2 = \varphi_{-1}(\sigma_1)\varphi_{-1}(\sigma_2)\\ &s_3 = \quad s_4 = \tfrac{1}{2}[\varphi_1(\sigma_1)\varphi_{-1}(\sigma_2) + \varphi_1(\sigma_2)\varphi_{-1}(\sigma_1)]\\ &a_1 = 0 \qquad\qquad\qquad\qquad a_2 = 0\\ &a_3 = -a_4 = \tfrac{1}{2}[\varphi_1(\sigma_1)\varphi_{-1}(\sigma_2) - \varphi_1(\sigma_2)\varphi_{-1}(\sigma_1)].\end{aligned}\right\} \qquad (396)$$

If we take (392′) into account and recall that

$$\int \mathcal{S}^*\mathcal{L}\mathcal{A}\, dS = \int \mathcal{A}^*\mathcal{L}\mathcal{S}\, dS = 0 \tag{397}$$

(because of the antisymmetric nature of the integrand), we can write (395) as

$$L^{jl} = 2I_S \sum_{\sigma_1\sigma_2} a_j^* a_l + 2I_A \sum_{\sigma_1\sigma_2} s_j^* s_l, \tag{398}$$

where we have put

$$I_S = \int \mathcal{S}^*\mathcal{L}\mathcal{S}\, dS, \qquad I_A = \int \mathcal{A}^*\mathcal{L}\mathcal{A}\, dS. \tag{399}$$

The two double summations for the various pairs $(j,\ l)$ are calculated by use of (391) and (389), and we find the following result for the matrix of the L^{jl}:

$$\left\{\begin{matrix} 2I_A & 0 & 0 & 0 \\ 0 & 2I_A & 0 & 0 \\ 0 & 0 & (I_A + I_S) & (I_A - I_S) \\ 0 & 0 & (I_A - I_S) & (I_A + I_S) \end{matrix}\right\}. \tag{399′}$$

The secular equation which yields the ϵ^i may then be written as

$$(2I_A - \epsilon)^2 \begin{vmatrix} I_A + I_S - \epsilon & I_A - I_S \\ I_A - I_S & I_A + I_S - \epsilon \end{vmatrix} = 0,$$

and its roots are

$$\epsilon^1 = \epsilon^2 = \epsilon^3 = 2I_A, \qquad \epsilon^4 = 2I_S. \tag{400}$$

Substituting (399) and (400) into the systems (394) we find that for $i = 1, 2, 3$, the first two equations are identically satisfied, and the other two are equivalent (if we suppose, and we shall,[10] that $I_S \neq I_A$) to $c^{i3} = c^{i4}$, while c^{i1}, c^{i2}, c^{i3} remain arbitrary. However, for $i = 4$ we find from the first two equations that $c^{41} = c^{42} = 0$, and from the other two that $c^{43} = -c^{44}$. Finally we have

$$\text{for } i = 1, 2, 3{:}\quad \Psi_a^i = c^{i1}\psi_a^1 + c^{i2}\psi_a^2 + c^{i3}(\psi_a^3 + \psi_a^4);$$
$$\text{for } i = 4{:}\quad \Psi_a^4 = c^{43}(\psi_a^3 - \psi_a^4).$$

The coefficients c contained in these formulas remain arbitrary, except for the normalization conditions. If we take (392′) and

---

[10] The case $I_S = I_A$ corresponds to the case in which the exchange integral vanishes. (See footnote on page 461.)

(396) into account and put

$$\begin{aligned} \bar{s}_i &= c^{i1}s_1 + c^{i2}s_2 + 2c^{i3}s_3, \\ \bar{a} &= 2c^{43}a_3, \end{aligned}$$

the preceding expressions are written:

$$\left.\begin{aligned} \psi_a^i &= \sqrt{2}\,\mathfrak{A}\bar{s}_i, \\ \psi_a^4 &= \sqrt{2}\,\mathfrak{S}\bar{a}. \end{aligned}\right\} \qquad (i = 1, 2, 3) \qquad (401)$$

(Note that in the case $n_1 = n_2$, we have $\mathfrak{A} = 0$, and hence the states corresponding to $i = 1, 2, 3$ are absent.)

The first three eigenfunctions correspond (in our approximation) to the eigenvalue

$$E_{n_1} + E_{n_2} + 2I_A, \qquad (402)$$

and the fourth corresponds to the eigenvalue

$$E_{n_1} + E_{n_2} + 2I_S. \qquad (402')$$

In general, therefore, the original level $E_{n_1} + E_{n_2}$ splits into a triplet level[11] (402) and a singlet level (402′). For $n_1 = n_2$ the triplet level is absent. As may be seen from (401), the eigenfunctions for the triplet are obtained by multiplying an antisymmetric Schrödinger eigenfunction by a symmetric spin eigenfunction, whereas the singlet ψ is the product of a symmetric Schrödinger eigenfunction and an antisymmetric spin function. The separation between the singlet and triplet levels is given by $2(I_S - I_A)$, or, as is easily found with the aid of (399) and (390), by

$$\int \Psi_{n_1}^*(x_1)\Psi_{n_2}^*(x_2)\mathfrak{L}\Psi_{n_1}(x_2)\Psi_{n_2}(x_1)\, dS. \qquad (402'')$$

This is exactly the *exchange integral* defined in the preceding section, computed, however, by accounting only for the eigenfunctions in x, y, z without the spin factors.

This result finds a meaningful expression in the vector model, in which two spins oriented in the same direction, or "parallel" spins, are made to correspond to a symmetric spin eigenfunction, whereas

---

[11] If we take the forces due to the spins into account, this triplet level will in turn split into three singlet levels, except for the case where the position eigenfunction $\mathfrak{A}$ represents a state with zero azimuthal quantum number ("s-state"), in which the perturbation energy due to spin is zero, and the three levels of the triplet stay together.

an antisymmetric spin eigenfunction corresponds to spins in opposite directions, or "antiparallel" spins.[12] It may then be said that when Russell-Saunders coupling holds, it is legitimate to use the ordinary Schrödinger theory rather than the Dirac theory, with the following limitation (which expresses the Pauli principle) added: the eigenfunctions must be antisymmetric *in the position coordinates only* if the spins are parallel and must be symmetric if the spins are antiparallel. In the first case we have a triplet level (which the vector model interprets in terms of the three possible orientations which the total spin may have with respect to the orbital magnetic field), whereas in the second case (the resultant spin being zero) we can have only a singlet level. If then the two electrons are in "the same orbit," that is, if their orbital quantum numbers are equal (symbolically: $n_1 = n_2$), their spins cannot align themselves parallel, because in that case the two electrons would have all their four quantum numbers equal—a condition prohibited by the Pauli principle. Hence for $n_1 = n_2$ the triplet is absent, and only the singlet exists.

A characteristic trait of this quantum theory of "exchange" is that the energy difference between the singlet and corresponding triplet level exists, as we have seen, even when the magnetic actions upon the spin are neglected, whereas according to the vector model the two states would differ only in their opposite spin orientations. Hence their energy difference would be due only to the interaction between their magnetic moments. The reason is that in the preceding theory the spins, although not occurring in the expression of

---

[12] Consider the operator representing the square of the total spin, which is

$$\sigma^2 = (\sigma_x^{(1)} + \sigma_x^{(2)})^2 + (\sigma_y^{(1)} + \sigma_y^{(2)})^2 + (\sigma_z^{(1)} + \sigma_z^{(2)})^2,$$

where $\sigma_x^{(1)}$, $\sigma_y^{(1)}$, $\sigma_z^{(1)}$ are the operators corresponding to the spin components of the first electron (formed according to §45) and $\sigma_x^{(2)}$, $\sigma_y^{(2)}$, $\sigma_z^{(2)}$ are those for the second electron. Applying σ^2 to a function (of the spin variables) which is antisymmetric (a) or symmetric (s), we find, respectively,

$$\sigma^2 a = 0, \qquad \sigma^2 s = 8s.$$

This means, according to the fundamental principle of quantum mechanics, that a measurement of the total spin σ would give zero in the first case, $\sqrt{8}$ in the second case. (The vector model would give: $\sigma = 0$ for antiparallel spins, $\sigma = 2$ for parallel spins. The numerical deviation of the last value from $\sqrt{8}$ is due to the already noted imperfection of the vector model; see, for instance, §57 of Part II).

the Hamiltonian, enter indirectly through the application of the Pauli principle, since they compel us to select the symmetric or antisymmetric (position) eigenfunction according to whether the spins are antiparallel or parallel. It can be shown[13] that this selection is formally equivalent to assuming a very strong interaction energy between the spins, proportional to the cosine of the angle between them. Thus we see justification of the success which this modelistic hypothesis had in the interpretation of multiplet spectra before the advent of quantum mechanics.

If we were to develop the calculations of this section by taking the *symmetric* functions ψ_s rather than the ψ_a, we should find, by an identical procedure, that with fixed n_1 and n_2 there is a triplet of eigenfunctions of the form $\psi_s^i = \sqrt{2}\,\mathcal{S}\bar{s}_i$ $(i = 1, 2, 3)$, and a singlet eigenfunction of the form $\psi_s^4 = \sqrt{2}\,\mathcal{A}\bar{a}$. That is to say, under this hypothesis the triplet levels would correspond to parallel spins; the singlet levels to antiparallel spins. Since $\mathcal{A} = 0$ for $n_1 = n_2$, the singlet state rather than the triplet state would be absent. As will be pointed out in the following section, the experimental results on helium and similar spectra contradict this hypothesis, thus fully confirming the antisymmetry principle.

The preceding considerations have led us to divide the antisymmetric states (with respect to an exchange of all the coordinates) into two classes, according to whether they are *symmetric* or *antisymmetric* with respect to an exchange of the *position coordinates only*. The spectral terms of the first class constitute the "triplet system"; those of the second class, the "singlet system." We shall now prove a theorem analogous to that shown in §62, namely, that the transitions between the two classes of states are forbidden, or that the two systems of spectral terms do not combine with each other. However, though the theorem of §62 is rigorously true, the theorem we are about to prove holds only in the approximation of this section, where the spin forces are neglected. Hence these transitions must not be considered to be absolutely impossible but only highly improbable. In fact, as we shall see below in connection with the helium spectrum, some of the spectral lines corresponding to these transitions have been observed, but with very weak intensity. The proof is entirely analogous to that of §62. Since the operator $\mathfrak{H}$ does not involve the spin variables in the approximation

[13] No. 6 of the Bibliography, page 226.

considered here, it is necessarily symmetric with respect to the position coordinates. Hence from (364) it follows that if $\psi(t)$ is symmetric (or antisymmetric) with respect to the position coordinates, $d\psi$ will still have the same property, and hence also $\psi(t + dt)$.

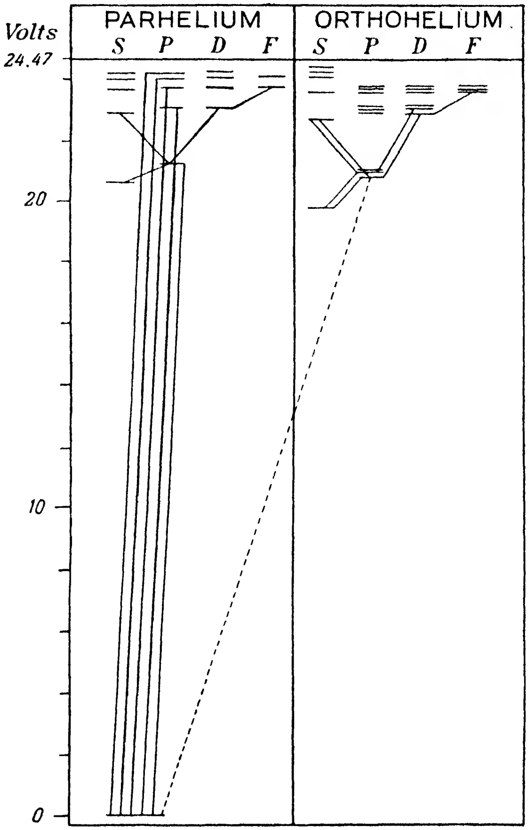

Fig. 48.

67. The spectrum of the helium atom and analogous spectra. The above results find one of their most important application in the case of the helium atom. Long before the rise of quantum mechanics, it was observed that the spectral lines of (non-ionized) helium could be divided into two classes, for each of which a system of separate terms was found (see Fig. 48). This discovery led to the belief that there were two distinct species of atoms, to which the names *parhelium* and *orthohelium* were given.[14] The lines of parhelium are all singlets, whereas those of orthohelium when examined with instruments of high resolving power generally reveal a fine-structure, characteristic of a system of triplets.[15] The ground level (which lies considerably lower than the others) belongs to the singlet system and corresponds to a 1*s*-state. The corresponding term in the triplet system is missing. Later an intercombination line was found, corresponding to a transition from one of the triplet levels to one of the singlet levels (see the broken line in Fig. 48), but its rather weak intensity indicates that such transitions occur with very small probability.

These experimental results are in excellent agreement with the theory presented in the preceding sections: the atoms of parhelium are those with antiparallel spins (that is, those whose ψ is antisymmetric in the position coordinates), whereas the atoms of orthohelium are those with parallel spins (or with a ψ symmetric in the position coordinates). In the ground state, both electrons are naturally in the lowest state, which is the 1*s*-state. They must therefore possess antiparallel spins in order to obey the Pauli principle, and hence this state belongs to the singlet series and has no analogue among the triplets. (This conclusion represents the confirmation of the postulate mentioned in the last section, that the ψ which actually occur are the antisymmetric ones, rather than the symmetric ones.) In the excited states, one of the electrons continues to occupy the 1*s* orbit, while the other electron moves to higher orbits. Hence their spins may be oriented both parallel and antiparallel.

We shall now outline the quantitative comparison between theory and experiment. This constitutes one of the most remark-

---

[14] The reason for these names lies in the former belief that the orbits of the two electrons were coplanar in parhelium and perpendicular in orthohelium. The names have persisted, although the model (which has never led to quantitatively satisfactory results) has been abandoned.

[15] This does not mean that the terms of orthohelium are all triplets but rather that they are triplets with the exception of the *S* terms, which are always singlets (in this connection see a volume on spectroscopy, or No. 27 of the Bibliography). We might add that the triplets of orthohelium are very narrow, so that many of them cannot be resolved or can be resolved only partially, so that they look like doublets. Their triplet nature is confirmed, however, by the analogous spectra (Li^+, Be^{++}, B^{+++}, C^{++++}), in which the separations are generally much larger.

able successes of quantum mechanics, especially because all previous attempts at theoretical interpretation on the basis of the Sommerfeld theory and others failed completely. Quantum mechanics contains the necessary elements for calculating not only the numerical values of the separate spectral terms but also all other physical constants of the helium atom, such as diamagnetic susceptibility, dielectric constant, van der Waals' constant, and boiling point. However, these calculations are quite laborious, since methods of successive approximations must necessarily be applied, and have given rise to numerous works of a mathematical nature.[16] We shall confine ourselves to enumerating some examples of the results. The quantity which has been calculated with the greatest care is the numerical value of the ground term (in other words, the ionization potential). These calculations have given

$$198{,}308 \text{ cm}^{-1}$$

as compared with the experimental result:

$$198{,}298 \pm 8 \text{ cm}^{-1};$$

the relative difference is 1 part in 20,000. The ionization potential turns out to be about 24.47 volts. For the excited states, the numerical calculation is more difficult and has not been pushed to such a high degree of accuracy, but it is always in satisfactory agreement with experiment. Concerning the effects of an electric or magnetic field, we restrict ourselves to citing the following results:

	Calculated	*Observed*
Magnetic susceptibility.......	-1.87×10^{-6}	-1.88×10^{-6}
Dielectric constant...........	1.000071	1.000074

As is to be expected, spectra analogous to the helium atom are emitted by the ions Li^+, Be^{++}, B^{+++}, C^{++++}, which, like the helium atom, have two electrons and differ from it only in the nucleus. Hence the formulas of He also apply to them after the numerical values have been changed. In these cases too, the agreement with experiment is excellent, as is shown by the following tabulation of

---

[16] An over-all presentation of these works, with bibliographical references, can be found in No. 1 of the Bibliography, from which the numerical data quoted here are taken.

the values of the ground terms in cm^{-1}:

	Calculated	*Observed*
Li^+	609,985	610,090 ± 100
Be^{++}	1,241,222	1,241,350 ± 200
B^{+++}	2,091,770	2,092,000 ± 300
C^{++++}	3,161,770	3,161,900 ± 800

Finally, it should be added that the atoms of the second column of the periodic system, or alkaline earths, present a certain analogy to the helium atom as far as the criterion in §59 of Part II is concerned, when the two valence electrons are considered to be subject to the attraction of the nucleus diminished by the "screening" action of the remaining $Z - 2$ electrons (which form completed shells, and hence may to a good approximation be replaced by a static space charge; see footnote on page 253). Therefore the problem differs from the helium case, because the field in which the two electrons move is not Newtonian. In this case, of course, the theory may not be made to yield very exact quantitative predictions, but its qualitative conclusions are confirmed by experiment. In fact, we still obtain a singlet system and a triplet system in these cases (the latter with considerably larger separations than in helium). The intercombinations between singlets and triplets are somewhat less improbable, however, as predicted by the theory and confirmed by observation. For this and other spectroscopic applications of quantum mechanics, see a text on spectroscopy. (See footnote, page 252.)

68. Note concerning other applications. The theory of systems with equal particles presented in this chapter—in which the quantum mechanical concept of *resonance* or *exchange* explained in §65 is characteristic—has also been successfully applied to the study of the collisions of slow electrons with hydrogen or helium atoms.[17] In the latter phenomenon the exchange between the colliding electron and those of the atom influences considerably the angular distribution of the electrons after the collision, and this influence is revealed in the experimental curves. Another important application is the explanation, due to Heisenberg,[18] of the origin of the

[17] See Mott and Massey, *The Theory of Atomic Collisions*, Oxford: Clarendon Press, 1949; or G. Wentzel, *Wellenmechanik der Stoss- und Strahlungsvorgänge*, in Geiger and Scheel, *Handbuch der Physik*, XXIV$_1$, 2nd edition, Berlin: J. Springer, 1933.

[18] *Zeits. f. Physik* **49**, 619 (1928).

so-called Weiss "molecular field," which is necessary to explain ferromagnetism but for which no satisfactory justification was found until Heisenberg's explanation. Most important among the applications of the exchange concept is the theory of the hydrogen molecule H_2, due to Heitler and London, who for the first time gave a satisfactory explanation of the force holding the two hydrogen atoms together. This force is essentially due to the afore-mentioned phenomenon of exchange or resonance between the two electrons and is therefore called *exchange force.*[19] We confine ourselves here to pointing out that these "exchange forces" generally play an important part in the interactions between neighboring atoms constituting a molecule or a crystal lattice, and presumably also in the interactions between the particles constituting the atomic nucleus. Furthermore, the physical theory of valence, introduced by Heitler and London for the case of the H_2 molecule, has been extended, at least qualitatively, to more complex molecules with considerable success, and seems capable of further development.

VALUES OF SOME FUNDAMENTAL CONSTANTS*

(CGS Units)

Velocity of light	$c =$	2.99776×10^{10}
Planck's constant	$h =$	6.6237×10^{-27}
Avogadro's number†	$N =$	6.0235×10^{23}
Faraday (in e.m.u.)†	$F =$	9649.6
Electronic charge:		
In e.s.u	$e =$	4.8024×10^{-10}
In e.m.u	$e/c =$	1.60199×10^{-20}
Specific electronic charge:		
In e.s.u	$e/m =$	5.2741×10^{17}
In e.m.u	$e/mc =$	1.75936×10^{7}
Electronic (rest) mass	$m =$	9.1055×10^{-28}
Ergs corresponding to 1 electron-volt		1.60199×10^{-12}
Ratio of proton mass to electron mass	$M_p/m =$	1836.57
Proton mass	$M_p =$	1.67229×10^{-24}
Bohr magneton	$\mu_0 =$	0.92736×10^{-20}
Fine-structure constant‡	$\alpha = \frac{2\pi e^2}{hc} =$	0.72969×10^{-3}

* All values are taken from J. W. M. DuMond and E. R. Cohen, *Rev. Mod. Phys.* **20,** 82 (1948) and *ibid.,* **21,** 651 (1949).

† The constants depending on the gram-molecular weight are here given on the *chemical* scale (see §1, Part I, footnote 2).

‡ R. T. Birge, *Phys. Rev.* **79,** 193 (1950).

---

[19] For a presentation of this theory we refer the reader to the volume *Molecole e cristalli* ("Molecules and Crystals") of the Italian physics series.

Bibliography*

(A) WORKS ON QUANTUM MECHANICS AND WAVE MECHANICS

1. Bethe, H., "Quantenmechanik der Ein- und Zwei-Elektronenprobleme," *Handbuch der Physik* XXIV, Part I, Chap. 3, 2nd ed., Berlin: J. Springer, 1933. (See No. 14 below.)

2. Bloch, E., *L'ancienne et la nouvelle théorie des quanta,* Paris, 1930.

3. Condon, E. U., and Morse, P. M., *Quantum Mechanics,* New York: McGraw-Hill, 1929.

4. de Broglie, L., *La Théorie de la quantification dans la nouvelle mécanique,* Paris, 1932.

5. ———, *Introduction à l'étude de la mécanique ondulatoire,* Paris, 1931.

6. ✓ Dirac, P. A. M., *The Principles of Quantum Mechanics,* 3rd ed., Oxford: Clarendon Press, 1947.

7. Frenkel, J., *Wave Mechanics, Elementary Theory,* Oxford: Clarendon Press, 1932.

8. ———, *Wave Mechanics, Advanced General Theory,* Oxford: Clarendon Press, 1934.

9. Haas, A., *Materiewellen und Quantenmechanik,* Leipzig, 1928.

9a. ———, *Wave Mechanics and the New Quantum Theory,* London: Constable, 1928.

10. Heisenberg, W., *Die physikalischen Prinzipien der Quantentheorie,* Leipzig, 1930.

10a. ———, *The Physical Principles of the Quantum Theory,* Chicago: University Press, 1930.

10b. Heitler, W., *Elementary Wave Mechanics,* Oxford: Clarendon Press, 1945.

10c. Kemble, E. C., *Fundamental Principles of Quantum Mechanics,* New York: McGraw-Hill, 1937.

10d. Kramers, H. A., "Die Grundlagen der Quantentheorie" and "Quantentheorie des Elektrons und der Strahlung," *Theorien des Aufbaues der Materie,* Vol. I, Leipzig, 1938.

10e. Jordan, P., *Anschauliche Quantentheorie,* Berlin: J. Springer, 1936.

11. March, A., *Die Grundlagen der Quantenmechanik,* Leipzig, 1931.

---

* This list is limited to just a few of the numerous works related to the subjects treated in this volume. More complete bibliographical references may be found in Nos. 2, 8, 20, and 21b.

12. Mott, N. F., *An Outline of Wave Mechanics,* Cambridge: University Press, 1930.

12a. ———, and Sneddon, I. N., *Wave Mechanics and its Applications,* Oxford: Clarendon Press, 1948.

13. von Neumann, J., *Mathematische Grundlagen der Quantenmechanik,* Leipzig, 1931.

14. Pauli, W., "Die allgemeinen Prinzipien der Wellenmechanik," *Handbuch der Physik* **XXIV**, Part I, Chap. 2, 2nd ed., Berlin: J. Springer, 1933. (See No. 1 above.)

14a. Pauling, L., and Wilson, E. B., *Introduction to Quantum Mechanics,* New York: McGraw-Hill, 1935.

15. Persico, E., *Lezioni di meccanica ondulatoria,* Padua, 1935.

15a. Rojansky, V. B., *Introductory Quantum Mechanics,* New York: Prentice-Hall, 1938.

16. Ruark, A. E., and Urey, H. C., *Atoms, Molecules, and Quanta,* New York: McGraw-Hill, 1930.

16a. Schiff, L. I., *Quantum Mechanics,* New York: McGraw-Hill, 1949.

17. Schrödinger, E., *Abhandlungen zur Wellenmechanik,* Leipzig, 1928.

17a. ———, *Collected Papers on Wave Mechanics,* London: Blackie, 1928.

17b. ———, *Mémoires sur la mécanique ondulatoire,* Paris, 1933.

18. Sommerfeld, A., *Atombau und Spektrallinien, Wellenmechanischer Ergänzungsband,* Braunschweig: Vieweg, 1929.

19. ———, *Wave Mechanics,* London: Methuen, 1930.

20. Weyl, H., *Gruppentheorie und Quantenmechanik,* Leipzig, 1931.

20a. ———, *The Theory of Groups and Quantum Mechanics,* London: Methuen, 1931.

(B) WORKS TO BE CONSULTED FOR SPECIAL ARGUMENTS

20b. Bethe, H. A., *Elementary Nuclear Theory,* New York: Wiley, 1947.

21. Bohr, N., *La Théorie atomique et la description des phénomènes,* Paris, 1932.

21a. ———, *Atomic Theory and the Description of Nature,* Cambridge: University Press, 1934.

21b. Born, M., *Atomic Physics,* New York: Hafner, 1946.

22. Brillouin, L., *Notions de mécanique ondulatoire; les méthodes d'approximation,* Paris, 1932.

23. Brunetti, R., *L'atomo e le sue radiazioni,* Bologna, 1932.

24. Castelfranchi, G., *Fisica moderna,* Milano, 1931.

24a. ———, *Recent Advances in Atomic Physics,* 2 vol., Philadelphia: Blakiston, 1932.

25. Courant, R., and Hilbert, D., *Methoden der mathematischen Physik,* 2nd ed., Berlin: J. Springer, 1931.

26. de Broglie, L., *L'Electron magnétique,* Paris, 1934.

26a. ———, *Matter and Light,* New York: Norton, 1939.

27. Fermi, E., *Introduzione alla fisica atomica,* Bologna, 1928.

27a. Finkelnburg, W., *Atomic Physics,* New York: McGraw-Hill, 1950.

28. Frank, J., and Jordan, P., *Anregung von Quantensprüngen durch Stösse,* Leipzig, 1926.

29. Gamow, G., *Constitution of Atomic Nuclei and Radioactivity,* Oxford: Clarendon Press, 1932.

29a. ———, and Critchfield, C. L., *Theory of Atomic Nucleus and Nuclear Energy-Sources,* Oxford: Clarendon Press, 1949.

29b. Heitler, W., *Quantum Theory of Radiation,* 2nd ed., Oxford: Clarendon Press, 1944.

30. Mark, H., and Wierl, R., *Die experimentellen und theoretischen Grundlagen der Elektronenbeugung,* Berlin: J. Springer, 1931.

31. Millikan, R. A., *Electrons (+ and −), Protons, Photons, Neutrons, and Cosmic Rays,* Cambridge: University Press, 1935.

31a. Mott, N. F., and Massey, H. S. W., *Theory of Atomic Collisions,* 2nd ed., Oxford: Clarendon Press, 1949.

31b. Oldenberg, O., *Introduction to Atomic Physics,* New York: McGraw-Hill, 1949.

32. Persico, E., *Ottica,* Milano, 1932.

32a. ———, *Introduzione alla Fisica Matematica,* Bologna, 1943.

32b. ———, *Gli atomi e la loro energia,* Bologna, 1948.

33. Pincherle, S., and Amaldi, U., *Le operazioni distributive e le loro applicazioni all'analisi,* Bologna, 1901.

33a. Rasetti, F., *Elements of Nuclear Physics,* New York: Prentice-Hall, 1936.

33b. Richtmyer, F. K., and Kennard, E. H., *Introduction to Modern Physics,* 4th ed., New York: McGraw-Hill, 1947.

34. Frank, P., and von Mises, R., *Die Differential- und Integralgleichungen der Mechanik und Physik,* 2 vol., Braunschweig: Vieweg, 1927, 1930.

35. Tonelli, L., *Serie trigonometriche,* Bologna, 1928.

35a. Byerly, W. E., *Fourier Series and Spherical Harmonics,* New York: Ginn, 1893.

36. de Broglie, L., *La Mécanique ondulatoire des systèmes de corpuscules,* Paris, 1939.

37. Flügge, S., and Krebs, A., *Experimentelle Grundlagen der Wellenmechanik,* Dresden and Leipzig, 1936.

Index