Electrodynamics, Quantum

W. P. Healy
RMIT University

GLOSSARY

Anomalous magnetic moment Difference between the intrinsic magnetic moment of a charged spin-$\frac{1}{2}$ particle and that predicted by the single-particle Dirac theory.

Coherent state State of the quantized radiation field in which the average electric and magnetic fields and the average energy are the same as the corresponding quantities for a state of the classical electromagnetic field.

***C*, *P*, and *T* symmetries** Invariance of quantum electrodynamics under the operations of charge conjugation (or matter–antimatter interchange), parity (or left–right interchange), and time reversal, respectively.

Dirac equation Four-component relativistic quantum mechanical wave equation for a spin-$\frac{1}{2}$ particle.

Einstein's *A* and *B* coefficients Factors determining the rates of spontaneous emission, induced emission, and absorption of radiation by atoms.

Feynman diagram Pictorial representation of a process in quantum electrodynamics in which states of particles or atoms are depicted as lines and their interactions as vertices where two or more lines meet.

Gauge invariance Independence of a quantity of the choice of potentials used to represent the electromagnetic field.

Lamb shift Change in atomic energy levels (from the values predicted by the single-particle Dirac theory) caused by electromagnetic interactions, or the splitting of spectral lines due to this change.

Leptons Spin-$\frac{1}{2}$ particles subject to the weak and, if charged, the electromagnetic force, but not subject to the strong nuclear force, and including electrons, muons, tauons, neutrinos, and their antiparticles.

Maxwell's equations General fundamental equations for the electromagnetic field, summarizing the basic laws of electromagnetism.

Occupation-number state State of a quantized field (such as the Maxwell or Dirac field) that has a definite number of particles or quanta in each field mode.

Photon Particle or quantum of the electromagnetic field that travels at the speed of light, has no charge or rest mass, and has intrinsic spin 1.

Renormalization Elimination of unobservable mass and charge of bare particles in favor of observed mass and charge of physical particles.

***S*-matrix element** Probability amplitude for a scattering process in which the incoming and outgoing particles are specified by their momenta and polarization or spin states.

QUANTUM ELECTRODYNAMICS is the fundamental theory of electromagnetic radiation and its interaction with microscopic charged particles, particularly electrons and positrons. In its most accurate form, the theory combines the methods of quantum mechanics with the principles of special relativity; often, however, it is sufficient to treat the charged particles in nonrelativistic approximation. Each part of the complete dynamical system of radiation and charges displays a characteristic wave–particle duality. Thus, electrons behave in many circumstances as particles, but they can also exhibit wave properties such as interference and diffraction. Similarly, electromagnetic radiation, which was considered classically as a wave field, may have particle properties ascribed to it under suitable conditions (e.g., in scattering experiments). The particles or quanta associated with the electromagnetic field are called photons.

Quantum electrodynamics is a highly successful theory, despite certain mathematical and interpretational difficulties inherent in its formulation. Its success is due in part to the weakness of the coupling between the radiation and the charges, which makes possible a perturbative treatment of the interaction of the two parts of the system. The theory accounts for many phenomena, including the emission or absorption of radiation by atoms or molecules, the scattering of photons or electrons, and the creation or annihilation of electron–positron pairs. Its most famous predictions concern the electromagnetic shift of energy levels observed in atomic spectra and the anomalous magnetic moment of the electron; both of these predictions are in good agreement with experimental results. Quantum electrodynamics also includes the interaction of photons with muons and tauons (which differ from electrons only in mass) and their antiparticles. The validity of the theory has been tested in high-energy collision experiments involving these particles down to distances less than 10^{-16} cm.

I. INTRODUCTION

A. Early Theories of Light

Since quantum electrodynamics is the modern theory of electromagnetic radiation, including visible light, it is instructive to begin with a brief historical review of previous theories. The nature of light has long been a subject of interest to philosophers and scientists. In the fifth century B.C., Empedocles of Acragas held that light takes time to travel from one place to another but that we cannot perceive its motion. He knew that the moon shines by light reflected from the sun and was also aware of the cause of solar eclipses. Heron of Alexandria, who is thought to have lived in the first or second century A.D., discussed the rectilinear propagation properties of light. In his book *Catoptrica* he derived the law of reflection using a principle of minimal distance. The law of refraction was not formulated until 1621, when it was discovered experimentally by Snell. Snell's law was later derived theoretically from Fermat's celebrated principle of least time.

From about the middle of the seventeenth century to the end of the nineteenth century there were two competing, and mutually contradictory, theories of light. The wave theory was initiated by Hooke and Huygens following the first observations of interference and diffraction. Huygens enunciated a principle, based on the wave theory, from which he derived the laws of reflection and refraction. He also discovered the polarization properties of light. These properties, as well as the law of rectilinear propagation, were difficult to explain by the wave theory, which at that time dealt only with longitudinal waves in a hypothetical "aether," analogous to sound waves in air. These difficulties led Newton to propose a corpuscular theory, according to which light is emitted from luminous bodies in a stream of small particles or corpuscles. Newton's views inhibited any further advances in the wave theory until about the beginning of the nineteenth century. In the meantime, the fact that light has a finite speed was confirmed by Römer through observations of eclipses of the moons of Jupiter. This occurred in 1675, more than two millenia after the time of Empedocles. (The speed of light in empty space is denoted by c and is approximately 2.998×10^{10} cm/sec in cgs units.)

B. Classical Electrodynamics

The revival of the wave theory began with Young's interpretation of interference experiments. In particular, the destructive interference of two light beams at certain points in space seemed totally inexplicable on the corpuscular hypothesis but was readily accounted for by the wave theory. Young also suggested that light waves execute transverse rather than longitudinal vibrations, as this could then explain the observed polarization properties. The wave theory was further developed by Fresnel, who applied it to phenomena involving diffraction, interference of polarized light, and crystal optics. An important test of the theory was provided by the comparison of the speeds of light in media with different refractive indexes. According to the wave theory, light travels slower in an optically denser medium, but according to the corpuscular theory it travels faster. The results of experiments carried out in 1850 agreed with the predictions of the wave theory.

The wave theory was in a certain sense completed when Maxwell established his equations for the electromagnetic field and showed that they have solutions corresponding to transverse electromagnetic waves in which both the electric and magnetic induction field vectors oscillate perpendicularly to the direction of propagation. The speed of these waves in empty space could be calculated from constants (the permittivity and permeability of the vacuum) obtained by purely electric and magnetic measurements and was found to be the speed of light. This conclusion became the basis of the electromagnetic theory of light. It was subsequently found that the frequencies of visible light form only a small part of the complete spectrum of electromagnetic radiation, which also includes radio waves, microwaves, and infrared radiation on the low-frequency side, and ultraviolet radiation, X-rays, and gamma rays on the high-frequency side.

C. Photons

Despite the success of classical electromagnetic theory in dealing with the propagation, interference, and scattering of light, experiments carried out about the end of the nineteenth century and the beginning of the twentieth century led to the reintroduction of the corpuscular theory, though in a form different to that proposed by Newton. The departure from classical concepts began in 1900 when Planck published his law of black-body radiation. In this law the quantum of action h (approximately 6.626×10^{-27} erg sec), now known as Planck's constant, made its first appearance in physics. Planck's law for the variation with frequency of the energy in black-body radiation at a given temperature is closely related to the existence of discrete energy levels for the electromagnetic field, even though Planck, in his original derivation of the law, did not consider the field itself to be quantized. A black body is one that absorbs all the electromagnetic energy incident on it. It was shown by Kirchhoff in 1860 that when such a body is heated, the emitted radiation does not depend on the detailed composition of the body but only on its absolute temperature. Radiation confined in a state of thermal equilibrium in a cavity with perfectly reflecting walls behaves as black-body radiation. According to classical electromagnetic theory, the cavity radiation can undergo simple harmonic motion at a number of certain allowed or characteristic frequencies ν, the values of which depend on the shape and size of the enclosure. These so-called radiation oscillators may be quantized, as in ordinary quantum mechanics. Then for each nonnegative integer n, an oscillator with frequency ν has a nondegenerate stationary state with energy $nh\nu$ above the ground-state energy (see Fig. 1). The possible values of the energy at this frequency thus form a discrete set 0, $h\nu$, $2h\nu$, $3h\nu$, ... instead of a continuum. It can be shown that quantization of the oscillators in this way for all the allowed frequencies leads directly to Planck's law.

In 1905, Einstein made use of the idea of light quanta in order to explain the photoelectric effect and later applied it to the emission as well as the absorption of radiation by atoms. The light quantum hypothesis states not only that the energy of monochromatic radiation of frequency ν is made up of integral multiples of the quantum $h\nu$, but also that the momentum is made up of integral multiples of the quantum h/λ, where λ is the wavelength of the radiation (ν and λ are related by the equation $\nu\lambda = c$). This hypothesis contrasts sharply with the classical picture in which the energy and momentum are regarded as continuously variable. The existence of discrete light quanta, or photons, is not immediately evident on a macroscopic scale, however. Due to the smallness of Planck's constant, even in a weak electromagnetic field there is an enormous number of photons, provided the frequency is not too high. For example, black-body radiation at a temperature of 300 K (room temperature) contains about 5.5×10^8 photons/cm^3, most of which correspond to frequencies in the infrared part of the electromagnetic spectrum. At a temperature of 6000 K (roughly that at the surface of the sun), the bulk of the radiation has frequencies in the visible spectrum, and there are about 4.4×10^{12} photons/cm^3. (The total number of photons in black-body radiation is proportional to the cube of the absolute temperature.)

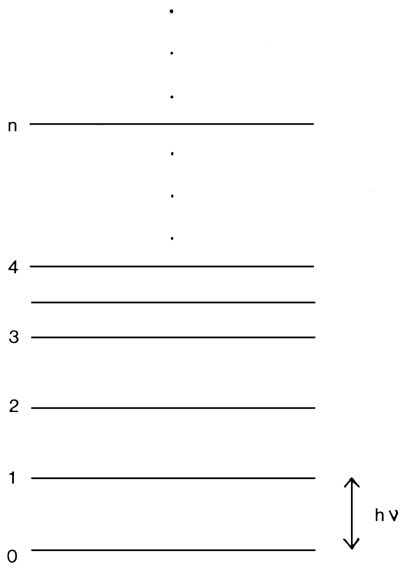


FIGURE 1 The horizontal lines represent the discrete energy levels of a quantized harmonic oscillator of frequency ν. The energy of the ground state is taken to be 0 and the equally spaced levels $0, h\nu, 2h\nu, \ldots, nh\nu, \ldots$ are labeled by the quantum numbers $0, 1, 2, \ldots, n, \ldots,$ respectively. Excitation of the radiation field at frequency ν to level $nh\nu$ corresponds to the addition of n photons, each with energy $h\nu$, to the field.

Individual photons manifest themselves only through their interaction with atomic systems. According to Einstein's treatment of the absorption and emission of radiation, for example, an atom in a stationary state can make a transition to a lower or a higher energy level accompanied

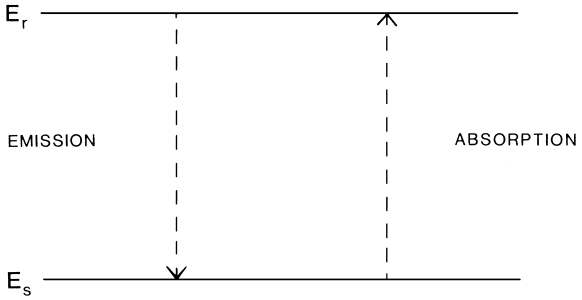


FIGURE 2 An atom can make a transition from a higher energy level E_r to a lower level E_s while emitting a photon of frequency ν, where $h\nu = E_r - E_s$. The emission may be spontaneous or induced by radiation. The atom can make the upward transition from level E_s to level E_r by absorbing a photon of frequency ν.

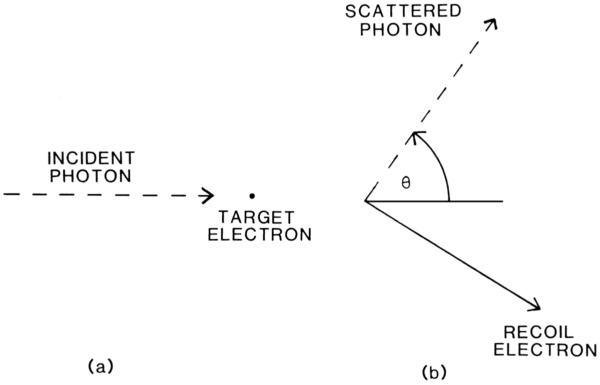


FIGURE 3 Photon and electron in the Compton effect (a) before collision and (b) after collision. The scattering angle θ is the angle between the initial and final directions of the photon.

by the creation or annihilation, respectively, of a photon. If the atomic energies are E_r and E_s, where $E_r > E_s$, then the energy $h\nu$ of the photon must equal the difference $E_r - E_s$ (see Fig. 2). This is called Bohr's frequency condition and is equivalent to the law of conservation of energy applied to the complete system of atom and radiation; any energy lost or gained by the atom is given up to or abstracted from the radiation field in the form of photons. It should be noted that the number of photons in the radiation field need not be constant—photons can be created or annihilated through the interaction of the field with atoms.

The scattering of X-rays by free electrons also furnishes direct evidence for the corpuscular properties of radiation. In 1922, Compton discovered that when X-rays of wavelength λ are incident on a graphite target, the scattered X-rays have intensity peaks at two wavelengths, λ and λ', where $\lambda' > \lambda$. The shift in wavelength given by $\Delta\lambda = \lambda' - \lambda$ is a function of the angle of scattering (i.e., the angle between the direction of the incident and scattered X-rays) but is independent of wavelength and the target material. The X-rays with unchanged wavelength λ were understood to have been elastically scattered by atoms, which suffer no appreciable recoil, and they could readily be accounted for on the basis of classical electrodynamics. The scattered X-rays with shifted wavelength λ', however, required a new interpretation. If it is assumed that the incident X-rays consist of photons, then these may collide with essentially free electrons in the target. In this case a photon gives up some of its energy $h\nu$ to an electron and is scattered with a lower frequency ν' and a longer wavelength λ', where $\nu'\lambda' = c$.

The wavelength shift $\Delta\lambda$ can be calculated as a function of scattering angle by using the laws of conservation of energy and momentum. By treating the problem relativistically and taking the electron to be at rest initially (see Fig. 3), one can easily show that $\Delta\lambda$ depends on the scattering angle θ alone through the formula

$$\Delta\lambda = \lambda_c(1 - \cos\theta),$$

where the constant λ_c is the Compton wavelength given by

$$\lambda_c = h/mc \simeq 2.43 \times 10^{-10}\ \text{cm}.$$

Here m is the rest mass of the electron (approximately 9.11×10^{-28} g). This formula was verified experimentally. The energy and distribution of the recoil electrons and scattered X-rays were also in accord with the predictions of the photon theory.

D. Quantum Electrodynamics

The use of the photon concept to explain certain phenomena does not imply a return to a naive classical particle view of light and other forms of electromagnetic radiation. A proper account must also be given of the wave properties of radiation, such as interference and diffraction. Indeed, the formulas for the energy and momentum of the photons—$h\nu$ and h/λ—are based on the assumption that the photons are associated with waves of definite frequency and wavelength. The nature of electromagnetic radiation is such that it appears, *under different experimental conditions*, sometimes to have particle properties and sometimes to have wave properties—these two aspects are said to be complementary. A single coherent theory that encompassed the dual nature of radiation, and with it settled the age-old controversy between the wave theory and the corpuscular theory, was made possible only by the development of quantum mechanics in the mid-1920s. In 1927, Dirac used the new methods of quantization, which had been successfully applied to atomic systems, to quantize the radiation field enclosed in a cavity, and was thus able to give a fully dynamical treatment of the emission and absorption of light by atoms. The beginning of quantum electrodynamics may be taken to date from this time.

The wave properties of radiation can be adequately described by using Maxwell's equations for the electromagnetic field, and these are retained as operator equations in the quantum theory of radiation. Suppose, for example, that the electromagnetic field has a node (i.e., a point where the field amplitudes always vanish) due to interference at P. Then an atom placed at P has, in so far as it can be regarded as a geometrical point, zero probability of absorbing a photon from the field. The field amplitudes are, however, subject to uncertainty relations, involving Planck's constant h, which are analogous to the Heisenberg uncertainty relations for the position and momentum of a particle in ordinary quantum mechanics. The origin of the uncertainty relations for the fields may be understood by considering a simple example.

Let $\bar{\mathcal{E}}$ denote the average value of a component of the electric field over a volume V and a time interval T. (Since a field component at a definite point in space and a definite instant of time appears an abstraction from physical reality, only such average values need be considered.) Now $\bar{\mathcal{E}}$ may be found by measuring the change produced by the field in the momentum of a charged test body occupying the volume during this time. Although the position and momentum of the test body are uncertain by amounts Δq and Δp that satisfy the Heisenberg uncertainty relation $\Delta q\ \Delta p \sim h$, it can be shown that this does not impair the accuracy of the field measurement, provided a sufficiently massive and highly charged body (which is therefore part of a macroscopic measuring instrument) is used. The charge Q on the body must be such that the product $Q\Delta q$ is large; $\bar{\mathcal{E}}$ can then be measured to any desired accuracy $\Delta\bar{\mathcal{E}}$. However, in the measurement of two average field strengths $\bar{\mathcal{E}}$ and $\bar{\mathcal{E}}'$ (taken over two space regions V and V' during two times intervals T and T', respectively), it may not be possible to make both $\Delta\bar{\mathcal{E}}$ and $\Delta\bar{\mathcal{E}}'$ as small as desired. If the separation distance L between V and V' (see Fig. 4) is such that most light signals emitted from V during the time interval T will reach V' during the time interval T', then the measurement of $\bar{\mathcal{E}}$ will influence that of $\bar{\mathcal{E}}'$ in a way that is to some extent unknown. The field produced by the test body used to measure $\bar{\mathcal{E}}$ is superimposed on $\bar{\mathcal{E}}'$ and cannot be fully subtracted out, as its value is somewhat uncertain (due to the uncertainty Δq in the position of the test body). This field, and hence $\Delta\bar{\mathcal{E}}'$, can indeed be made as small as desired by making the product $Q\Delta q$ sufficiently small, but then $\Delta\bar{\mathcal{E}}$ becomes relatively large. The experimental conditions for measurements of $\bar{\mathcal{E}}$ and $\bar{\mathcal{E}}'$ are complementary—those that serve to measure $\bar{\mathcal{E}}$ more precisely will meaeure $\bar{\mathcal{E}}'$ less precisely, and vice versa. The order of magnitude of the uncertainly product is given by

$$\Delta\bar{\mathcal{E}}\Delta\bar{\mathcal{E}}' \sim h/L^3T$$

and is independent of both Q and Δ_q. Thus, only for well-separated regions or over long intervals of time can both averages be measured with unlimited accuracy.

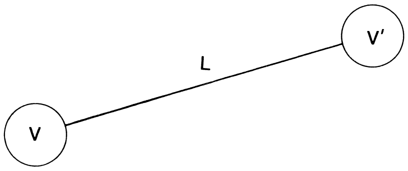


FIGURE 4 Regions V and V' in which the average electric fields $\bar{\mathcal{E}}$ and $\bar{\mathcal{E}}'$ are measured during time intervals T and T', respectively. The separation distance L is such that most light signals emitted from V during the interval T will reach V' during the interval T'.

E. Electrons and Positrons

Dirac's original radiation theory had to be modified to bring it into line with the special theory of relativity. This was true particularly of the treatment of the charged particles with which the electromagnetic field interacts. In 1928, Dirac had developed a one-particle relativistic wave equation for the electron that automatically accounted for the observed electron spin and predicted values for the fine structure of the energy levels of the hydrogen atom and of hydrogen-like ions that were in agreement with the experimental data of that time. The Dirac equation, however, also has extraneous solutions corresponding to negative-energy states. To eliminate these, Dirac introduced in 1930 the so-called hole theory, according to which most of the negative-energy states are occupied, each having one electron. Any unoccupied states, or holes, may be interpreted as particles with positive energy and positive charge. These particles were at first thought by Dirac to be protons, but were later identified as positrons, or antiparticles of electrons.

The experimental discovery of the positron by Anderson in 1932 lent support to Dirac's hole theory. Nevertheless, difficulties remained, such as the infinite (but unobservable) charge density associated with the "sea" of negative-energy electrons. These difficulties can be removed, however, by treating Dirac's one-particle wave equation for the electron as a field equation and subjecting it to a process of quantization, similar in some respects to the quantization of the classical electromagnetic field. This method, which is often referred to as second quantization, was applied to the Dirac equation by Heisenberg and others and resulted in the appearance of electrons and positrons, on an equal footing, as quanta of the Dirac field, just as photons appear as quanta of the Maxwell field. There are, however, some differences between the methods of second quantization used for the Dirac and Maxwell fields, which stem from the different characteristics of the associated particles or quanta. Photons have zero rest mass (but have nonzero momentum because they travel at the speed of light), are electrically neutral, have spin 1 (in units of $\hbar$, which is Planck's constant divided by 2π), and are bosons (i.e., any number of photons can occupy a given state). Electrons and positrons have the same nonzero rest mass, carry equal but oppositely signed charges (by convention, this is negative for the electron and positive for the positron), have spin $\frac{1}{2}$, and are fermions (i.e., not more than one electron or positron can occupy a given state). It was shown by Pauli that there is a connection between the spin of a particle and its so-called statistics—particles with integer spin are bosons and are not subject to the exclusion principle, whereas particles with half odd-integer spin are fermions and are subject to the exclusion principle. It is necessary for photons to be bosons in order that the quantized electromagnetic field may have a classical counterpart, which is realized in the limit of large photon occupation numbers. The quantized Dirac field, on the other hand, does not have a physically realizable classical limit.

F. Divergences and Renormalization

In quantum electrodynamics, as in classical electrodynamics, there are no known exact solutions to the equations for the complete dynamical system of radiation and charges. Indeed, from a purely mathematical viewpoint, the question of even the existence of such solutions is still an open one. Approximate solutions may be found by assuming that the coupling between the two parts of the system is weak and using perturbation theory. This is justified by the smallness of the fine-structure constant α, which gives a measure of the strength of the coupling:

$$\alpha = e^2/4\pi\hbar c \approx \tfrac{1}{137},$$

where e is the magnitude of the charge on the electron and rationalized cgs units are being used ($e \approx 1.355 \times 10^{-10}$ g$^{1/2}$cm$^{3/2}$/sec).

It was found in the 1930s and 1940s that the calculations for many processes, when taken beyond the first approximation, gave divergent results. Some divergences (the so-called infrared divergences) were due to deficiencies in the approximation method itself. Others (ultraviolet divergences) were associated with the problem of the structure and self-energy of the electron and other elementary particles. This problem had also arisen in classical electrodynamics, where the electron was assumed to have a structure-dependent electromagnetic contribution included in its inertial mass. In quantum electrodynamics, however, there occurred additional divergences of a radically different nature, due to effects that have no classical analogues. For example, the possibility of electron–positron pair creation gave rise to an infinite vacuum polarization in an external field and also implied an infinite self-energy for the photon.

The need to extract finite results from the formalism became acute when refinements in experimental technique revealed small discrepancies between the observed fine structure of the energy levels of atomic hydrogen and that given by Dirac's one-particle relativistic wave equation. These differences, whose existence had been suspected for some time, were measured accurately by Lamb and Retherford in 1947. In the same year, Kusch and Foley found that the value of the intrinsic magnetic moment of an electron in an atom also differs slightly from that predicted by the Dirac theory. For it to be shown that these discrepancies could be explained as radiative effects in quantum electrodynamics, it was necessary first to recognize that the mass and charge of the bare electrons and positrons that appear in the formalism cannot have their experimentally measured values. Since the electromagnetic field that accompanies an electron, for example, can never be "switched off," the inertia associated with this field contributes to the observed mass of the electron; the bare mechanical mass itself is unobservable. Similarly, an electromagnetic field is always accompanied by a current of electrons and positrons whose influence on the field contributed to the measured values of charges. The parameters of mass and charge, therefore, had to be renormalized to express the theory in terms of observable quantities. The results for the shift of energy levels (now known as the Lamb shift) and the anomalous magnetic moment of the electron then turned out to be finite and were, moreover, in good agreement with experimental results. The use of explicitly relativistic methods of calculation, developed by Tomonaga and Schwinger, was essential in avoiding possible ambiguities in this procedure. Further important contributions were made by Dyson, who showed that the renormalized theory gave finite results for interaction processes of arbitrary order, corresponding to arbitrary powers of the coupling constant e, and by Feynman, who introduced a diagrammatic representation of the mathematical expressions for these processes, which are often of considerable complexity.

The Feynman-diagram technique and Dyson's perturbation theory are now part of the standard formulation of quantum electrodynamics. This formulation and some of its applications will be outlined in Sections II and III. In this article only the electromagnetic interactions of electrons and positrons (or, more generally, of charged leptons, which include muons and tauons and their antiparticles) are considered. These particles also participate in the so-called weak interaction (and, of course, in the much weaker gravitational interaction). A unified theory of the electromagnetic and weak interactions has been developed in recent years. Many elementary processes, however, are dominated by electromagnetic effects, and these alone form the subject matter of quantum electrodynamics.

II. NONRELATIVISTIC QUANTUM ELECTRODYNAMICS

A. Approximations

Any treatment of the pure radiation field based on Maxwell's equations in empty space must satisfy the principles of the special theory of relativity, even though it might not be expressed in a form that makes this evident. Quantum electrodynamics has, however, a well-defined nonrelativistic limit in so far as the motion of the charged particles with which the electromagnetic field interacts is concerned. The nonrelativistic theory is of an approximate character, but it involves a much simpler mathematical formalism than that of its more exact relativistic counterpart. Moreover, it can be applied to a wide range of problems in physics and chemistry, particularly in the areas of atomic spectroscopy, intermolecular forces, laser physics, and quantum optics.

Nonrelativistic quantum electrodynamics provides an accurate description of phenomena when the following two conditions are satisfied:

1. The charged particles move at such slow speeds (in the inertial frame of a given observer) that their masses can be considered constant and equal to their rest masses. Since the relativistic mass of a particle with speed v and rest mass m is $m/\sqrt{(1-v^2/c^2)}$, this requires that $v/c \ll 1$. Now this inequality generally holds for the constituent particles of atoms under normal laboratory conditions. For example, the root-mean-square speed $\bar{v}$ (relative to the supposedly slowly moving nucleus) of the electron of a hydrogen-like ion in a state with principal quantum number n is $Ze^2/(2nh)$, where Ze is the nuclear charge. If $Z=1$ and $n=\ 1$ (the hydrogen atom in its ground state), then $\bar{v}/c$ equals the fine-structure constant α (approximately $\frac{1}{137}$) and the corresponding fractional increase in mass (over and above the rest mass) is only about 3 parts in 10^5. This ratio is larger for higher values of Z but smaller for higher values of n. The variation of mass with velocity is, therefore, expected to be appreciable only for the inner-shell electrons of the heavier elements.

2. The number of each type of charged particle (electron, proton, etc.) is conserved; that is, such particles are neither created nor destroyed in any process. This assumption imposes a restriction on the frequency ν of the radiation with which the particles may interact, since photons of sufficiently high energy are capable of creating particle–antiparticle pairs. This possibility requires an energy of order mc^2 (where m is the rest mass of the lightest charged particle, namely the electron) and will therefore be excluded if $\nu \ll \nu_c$, where ν_c is defined by $h\nu_c = mc^2$ and is about 10^{20} Hz. (Here ν_c is the frequency associated with the Compton wavelength of the electron given by $\lambda_c = h/mc$.) It follows that hard X-rays and high-energy gamma rays are to be omitted from consideration in this section.

B. An Assembly of Photons

The classical electromagnetic field in empty space is equivalent to an infinite number of one-dimensional simple harmonic oscillators. One oscillator is associated with each plane-wave component of the field, specified by its frequency ν, wave vector $\mathbf{k}$ (where $|\mathbf{k}| = 2\pi\nu/c$), and unit polarization vector $\hat{\mathbf{e}}$. The waves are transverse waves, which implies that the polarization vector is perpendicular to the direction of propagation $\hat{\mathbf{k}}$ (see Fig. 5). Hence, for each propagation direction, there are two independent polarization vectors $\hat{\mathbf{e}}^{(\lambda)}(\lambda = 1, 2)$. A radiation oscillator may therefore be labeled by the pair ($\mathbf{k}$, λ), which specifies the frequency, propagation direction, and polarization for the corresponding mode of the field.

A mathematical description of an assembly of noninteracting photons is obtained when each of the radiation oscillators is treated as a quantum mechanical system. This involves little more than the use of the matrix theory of the harmonic oscillator developed in elementary quantum mechanics but extended to cover the case of a set of independent oscillators. The result of this quantization of the electromagnetic field can be briefly summarized. States of the complete system are represented by vectors in a generalized (in fact, infinite-dimensional) vector space and

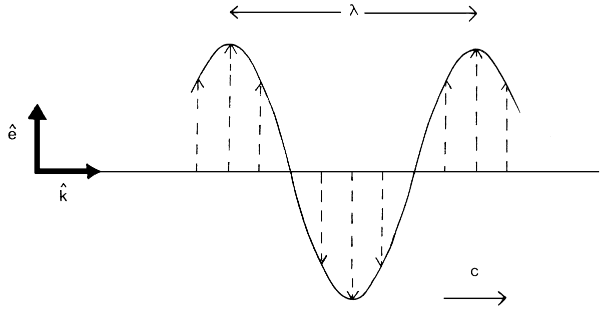


FIGURE 5 A linearly polarized electromagnetic wave of wavelength λ propagating in empty space with speed c in the direction $\hat{\mathbf{k}}$. The (real) unit polarization vector $\hat{\mathbf{e}}$ and $\hat{\mathbf{k}}$ together determine the plane of polarization, at any point of which the magnetic induction vector $\mathcal{B}$ (broken arrows) is parallel to $\hat{\mathbf{e}}$ and oscillates in simple harmonic motion of frequency ν, where $\nu\lambda = c$. The electric vector $\mathcal{E}$ (not shown) also oscillates with frequency ν and in phase with $\mathcal{B}$ but is perpendicular to the plane of polarization.

dynamical variables (such as energy and momentum) by linear operators, which act on the vectors to produce other vectors of the same kind. The vacuum state is that for which every oscillator has its lowest energy. It can be assumed, for convenience, that the energy of the vacuum state is zero. The so-called zero-point energy $1/2h\nu$ of an oscillator with frequency ν is therefore discarded, but this amounts merely to a shift in the datum point for measuring energies. (Nevertheless, changes in the zero-point energy can give rise to measurable forces, for example, between conducting plates. This is called the Casimir effect.)

If a radiation oscillator of mode $(\mathbf{k}, \lambda)$ is excited to its nth stationary state, with energy $nh\nu$, then this is taken to correspond physically to the presence of n photons, each with energy $h\nu$, for that mode of the field. An occupation-number state is one with a specified number of photons in each mode. The number $n_{\mathbf{k}\lambda}$ of photons in mode $(\mathbf{k}, \lambda)$ is then called the occupation number for that mode. Only a finite number of occupation numbers can be nonzero and the total numbers of photons is $\sum n_{\mathbf{k}\lambda}$ with the summation extending over occupied modes. Similarly, the total energy E of the photons in an occupation-number state is $\sum(n_{\mathbf{k}\lambda}h\nu)$. The vacuum state can be thought of as an occupation-number state for which every occupation number is zero.

The operator that represents the total energy of the radiation field is called the Hamiltonian operator and is denoted by H_{RAD}. It can be expressed in terms of photon annihilation and creation operators. The annihilation operator for mode $(\mathbf{k}, \lambda)$, when acting on an occupation-number state vector, reduces the number of photons for that mode by one, and when acting on the vacuum-state vector gives the zero vector. Similarly the creation operator increases the number of photons by one. (In the context of the elementary theory of the harmonic oscillator, these operators are usually called lowering and raising operators, respectively.)

A general state of the radiation field at time t is represented by a state vector Ψ of unit length that is a linear combination of the occupation-number state vectors, with coefficients $c(\ldots, n_{\mathbf{k}\lambda}, \ldots; t)$ depending on the occupation numbers and the time. The square of the magnitude of $c(\ldots, n_{\mathbf{k}\lambda}, \ldots; t)$ is the probability that if a measurement of the occupation numbers is carried out at time t, then these will be found to have precisely the values $\ldots, n_{\mathbf{k}\lambda}, \ldots$ So long as no measurements are made on the system, the time evolution of the state vector is governed by Schrödinger's equation:

$$i\hbar\,\frac{\partial\Psi}{\partial t} = H_{\text{RAD}}\Psi.$$

The (unit) length of the state vector does not change with time and so the dynamical behavior of the system may be said to correspond to a pure rotation in the generalized vector space. Indeed, the individual probabilities $|c|^2$ do not change with time, since the probability amplitudes c change only through a phase factor $\exp(-iEt/\hbar)$. This is consistent with the fact that for the free field, photons are neither created nor destroyed.

C. The Quantized Electromagnetic Field

While the treatment of the radiation field as an assembly of photons may seem to emphasize its corpuscular aspects, the wave properties are, nevertheless, also contained in the formalism. In particular, Maxwell's equations in empty space remain valid, although they now appear as operator equations rather than as equations for classical fields. Both the electric field $\mathcal{E}$ and the magnetic induction field $\mathcal{B}$ become operators that can be expressed as linear combinations of the photon annihilation and creation operators. If these expressions are inserted into the classical formula for the field energy, then the expansion of the Hamiltonian operator in terms of the annihilation and creation operators is recovered. Thus,

$$H_{\text{RAD}} = \frac{1}{2}\iiint(\mathcal{E}^2 + \mathcal{B}^2)\,dV.$$

(It is true that an infinite zero-point energy also appears. This energy may, however, be discarded, as were the zero-point energies of the individual oscillators.) Similarly, the classical expression

$$\frac{1}{c}\iiint \mathcal{E}\times\mathcal{B}\,dV$$

for the field momentum, obtained from Poynting's theorem, implies, when reinterpreted in terms of annihilation and creation operators, that a momentum $\hbar\mathbf{k}$ is to be ascribed to each photon of mode $(\mathbf{k}, \lambda)$. This is in agreement with Einstein's hypothesis, since $\hbar|\mathbf{k}| = h/\lambda$.

Another consequence of the quantization of the field is the occurrence of uncertainties or fluctuations that have no counterpart in the classical theory. In the vacuum state, for example, the mean value of the electric field is zero but its root-mean-square deviation from the mean, $\Delta\mathcal{E}$, is nonzero. The fluctuation $\Delta\mathcal{E}$ arises from the collective zero-point motions of the radiation oscillators and, if calculated for a nonzero volume with linear dimensions of order L and a nonzero time interval with length of order T, assumes a value whose order of magnitude is given by

$$\Delta\mathcal{E} \sim \begin{cases} \dfrac{\sqrt{\hbar c}}{L^2}, & \text{if} \quad L \geq cT \\[2ex] \dfrac{\sqrt{\hbar c}}{L(cT)}, & \text{if} \quad L \leq cT. \end{cases}$$

In any other of the occupation-number states, for which the mean values are also zero, the field fluctuations are of greater magnitude than those in the vacuum state. Now the occupation-number states resemble incoherent superpositions of classical plane-wave states, since they are associated with definite wave vectors and polarization vectors but do not have well-defined phases (in the classical sense), whereas a classical plane-wave field has a simple harmonic time dependence. There exist other states of the quantized field, however, called coherent or quasi-classical states, in which the phase is more well defined but the number of photons, and hence the energy and momentum, is less sharp than for the occupation-number states. To each state of the classical field there corresponds a unique coherent state of the quantized field such that (a) the mean values of the quantized field components are equal to the classical field components and (b) the mean value of the quantized energy is equal to the classical energy. The coherent states are also remarkable in the following respect: the field fluctuations for these states are exactly the same as those for the vacuum state.

The operators representing the components of the electric field $\mathcal{E}$ and the magnetic induction field $\mathcal{B}$ satisfy certain commutation relations, which may be derived from those for the photon annihilation and creation operators. Just as in ordinary quantum mechanics the commutation relation

$$qp - pq = i\hbar$$

between the position operator q and the momentum operator p of a particle leads to the Heisenberg uncertainty relation

$$\Delta q \, \Delta p \sim \hbar,$$

so the commutation relations between the components of $\mathcal{E}$ and $\mathcal{B}$ lead to uncertainty relations for the electromagnetic field strengths. These uncertainty relations are in agreement with the way in which the field strengths can, at least in principle, be measured by means of macroscopic test bodies. This was shown in detail by Bohr and Rosenfeld in 1933.

D. Interactions of Photons and Atoms

The quantized radiation field has so far been considered as a system by itself. A set of nonrelativistic charged particles, interacting through instantaneous Coulomb forces and also, perhaps, acted on by prescribed external static electric or magnetic fields, can also be considered as a system by itself, as in ordinary quantum mechanics. This system will, for convenience, be referred to as an atom, although it may really be a molecule, an ion, or a collection of atoms, molecules, or ions. It is assumed that there are N charged particles with masses $m_1, m_2, \ldots, m_N$; charges $e_1, e_2, \ldots, e_N$; position operators $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_N$; and momentum operators $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N$. The Hamiltonian operator for this system is given by

$$H_{\text{ATOM}} = \sum_{\alpha=1}^{N} \frac{1}{2m_\alpha} \mathbf{p}_\alpha^2 + U,$$

where the first term represents the kinetic energy and the second the potential energy. The potential energy U depends on the positions and momenta of the particles, their charges, and the external fields, if any are present.

It is often a good approximation to treat the nuclei as fixed and to regard the coordinates and momenta of the electrons alone as dynamical variables. This is possible because of the large mass of the protons and neutrons compared with that of the electrons (proton mass ≈ 1836 × electron mass). The fixed-nuclei approximation involves, among other things, the neglect of the recoil of the atoms which should accompany the absorption or emission of photons. The recoil velocity is, however, normally very small. For example, the speed imparted to a hydrogen atom by a photon with a frequency in the visible spectrum is of the order of 10^{-8} times the speed of light *in vacuo*. Such a speed results in only a very slight Doppler shift in the frequency of the emitted radiation. In the fixed-nuclei approximation, the Hamiltonian operator H_{ATOM} has, in general, a discrete set of energy levels $E_r, E_s, \ldots$ corresponding to bound states, as well as a continuous set of energy levels E corresponding to ionized states. Here $r, s, \ldots$ are shorthand notations for sets of quantum numbers sufficient to specify the states completely.

A state vector for the complete system consisting of the atom and the radiation field is obtained by multiplying a state vector for the field directly into a state vector for the atom. For example, there are states for which the photon occupation numbers have definite values and the atom is in a stationary state with a definite energy. The general state of the complete system is, at any instant, a superposition of such product states.

The Hamiltonian H for the complete system is not simply the sum of the radiation and atomic Hamiltonians given previously. This sum must be supplemented by an interaction term H_{INT}:

$$H = H_{\text{RAD}} + H_{\text{ATOM}} + H_{\text{INT}}.$$

The inclusion of the interaction term is essential if the operator equations of motion are to reproduce (a) Maxwell's equations for $\mathcal{E}$ and $\mathcal{B}$ with the charges and currents as sources and (b) the Lorentz-force law for the charged particles when they are acted on by $\mathcal{E}$ and $\mathcal{B}$, that is, the expected equations of motion for the interacting systems. The interaction Hamiltonian H_{INT} contains some operators that refer to the field and some that refer to the particles and, hence, is responsible for the coupling between the two parts of the complete system. In the absence of H_{INT}, the product vectors of the type mentioned above represent stationary states in which the photon occupation numbers are constant and the atom has a fixed energy. Due to the presence of H_{INT}, however, transitions between these states can occur, in which, for example, the atom loses or gains energy and the number of photons is correspondingly increased or decreased. The interaction Hamiltonian may be expressed as the sum of two parts, one proportional to e and the other to e^2:

$$H_{\text{INT}} = eH_1 + e^2 H_2,$$

where H_1 is linear and H_2 is quadratic in the photon annihilation and creation operators. As a consequence, these two terms give rise to processes in which the number of photons changes by one or two, respectively.

The use of the so-called Coulomb gauge is very convenient in nonrelativistic theory. In this gauge only the transverse electromagnetic field, which is a superposition of modes with transverse polarization vectors ($\hat{\mathbf{e}}^{(\lambda)} \cdot \hat{\mathbf{k}} = 0$), is quantized. The effect of the longitudinal field, responsible for the instantaneous Coulomb interaction between the charges, is treated as a potential as in ordinary quantum mechanics and is included in the expression for H_{ATOM}.

The time evolution of the complete system is governed by Schrödinger's equation:

$$i\hbar \frac{\partial \Psi}{\partial t} = H\Psi,$$

where now H is the total Hamiltonian and Ψ represents the state of both the field and the atom at time t. No exact solutions of this equation are known. Fortunately, however, H_{INT} is of order e and, hence, can be regarded as a small perturbation to the unperturbed Hamiltonian $H_{\text{RAD}} + H_{\text{ATOM}}$. Time-dependent perturbation theory can then be used to calculate approximately the probabilities for transitions between unperturbed states. The total energy is always exactly conserved in transitions between initial and final states. Since the perturbation is small, the unperturbed energy is approximately conserved in such transitions.

E. Applications

Applications of the theory to the emission and absorption of photons by atoms and the scattering of photons by free electrons will now be considered.

1. Spontaneous Emission—Einstein's *A* Coefficient

If initially (a) the atom is in an excited state r with energy E_r and (b) the radiation field is in the vacuum state, then there is a probability that after a time t a photon of mode ($\mathbf{k}$, λ) has been created and the atom has made a transition to a state s with lower energy E_s, where

$$h\nu \approx E_r - E_s.$$

Since there are no photons present initially, this process is known as spontaneous emission. It is represented graphically by the Feynman diagram in Fig. 6. Single-photon spontaneous emission involves, in the lowest order of perturbation theory, only that term in the interaction Hamiltonian that is proportional to e. Furthermore, the so-called dipole approximation can be used for optical or lower frequencies and bound states of atoms or small molecules, since then the wavelength of the emitted photon is much larger than the dimensions of the region in which the atomic wave functions differ significantly from zero. The emission probability can sometimes be expressed in terms of a constant transition rate (that is, a probability per unit time for the transition to occur) known as Einstein's A coefficient. The total transition rate for emission of the photon in any direction and with any polarization is given in dipole approximation by

$$A_s^r = (16\pi^3\nu^3/3hc^3)|\boldsymbol{\mu}^{rs}|^2,$$

where $\boldsymbol{\mu}^{rs}$ denotes the dipole transition moment, which can be calculated once the wave functions for the atomic states r and s are known. Thus, in dipole approximation, Einstein's A coefficient is proportional to the cube of the transition frequency and the square of the length of the dipole transition moment.

The reciprocal of A_s^r is the average lifetime of the upper state r with respect to the lower state s. For example, for optical transitions with a photon wavelength of order 5000 Å (1 Å $= 10^{-8}$ cm) and a dipole transition moment

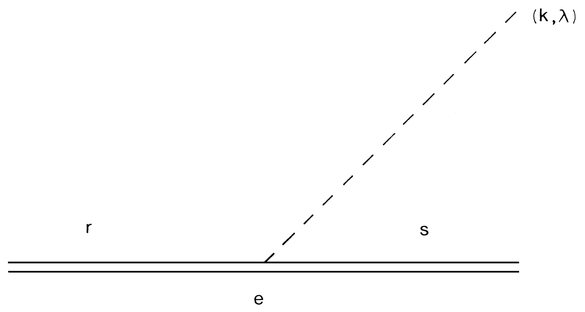


FIGURE 6 Feynman diagram for spontaneous emission. The left-hand and right-hand portions of the parallel horizontal lines represent the initial and final atomic states r and s, respectively. (Double lines are used to indicate that the electrons are not free but are bound to the atomic nucleus.) The dotted line represents the emitted photon of mode ($\mathbf{k}$, λ). This is created when the atom undergoes the transition $r \to s$. The vertex labeled e corresponds to the first-order term in the interaction Hamiltonian, which is responsible for this process in the lowest order of perturbation theory.

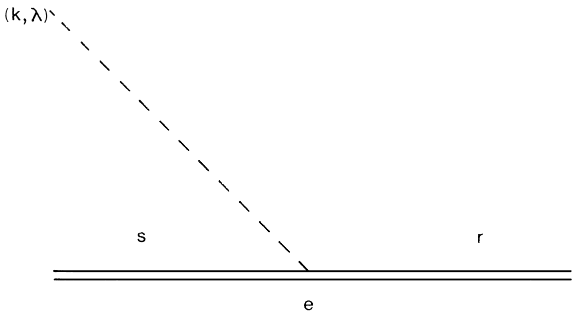


FIGURE 7 Feynman diagram for absorption. In the initial state (left of diagram), the atom has energy E_s and there is a photon of mode (**k**, λ) present, whereas in the final state (right of diagram), the atom has higher energy E_r and the photon has been annihilated.

of order ea_0 (where a_0 is the Bohr radius of hydrogen, approximately 0.53 Å), the lifetime is of order 10^{-8} sec.

The transition probability is proportional to the time t so long as t is large compared with the atomic period $1/\nu$ and small compared with the lifetime. Since, with the above assumptions, the period is of order 10^{-15} sec, there is indeed a range of values of t that satisfy both conditions. The detection of the emitted photons must take place at times t lying in this range, or else the emission rate is not approximately constant.

2. Absorption andhh Stimulated Emission—Einstein's *B* Coefficients

If the atom is initially at the lower level E_s, but there is radiation already present, it may make a transition to the hhigher level E_r by absorbing a photon with energy approximately equal to $E_r - E_s$. The Feynman diagram for absorption is shown in Fig. 7. The transition rate for this process is proportional to the photon occupation number $n_{\mathbf{k}\lambda}$ and hence to the intensity I (erg cm^{-3} Hz^{-1}) of the incident radiation in the spectral region from which the photon is absorbed. If the atom is bathed in isotropic unpolarized radiation (so that I is independent of $\hat{\mathbf{k}}$ and λ), the total absorption rate is $B_r^s I$ where B_r^s is Einstein's B coefficient for absorption, given in dipole approximation as

$$B_r^s = (2\pi^2/3h^2)|\boldsymbol{\mu}^{rs}|^2.$$

The upper limit on the time for the validity of this transition rate is now much less than the reciprocal of $B_r^s I$. Times less than this upper limit but much greater than the period can be found, provided the intensity of the radiation is not too high.

For an atom initially at the upper level E_r with radiation present as before, there is a probability for a transition to the lower level E_s accompanied by the emission of a photon with the same characteristics as some of those in the incident beam. This emission may, of course, occur spontaneously, that is, even when all the photon occupation numbers are zero. There is in addition, however, emission that is stimulated or induced by the incident radiation, at a rate proportional to its intensity. For isotropic unpolarized radiation, the stimulated emission rate is $B_s^r I$, where the B coefficient for emission $r \rightarrow s$ is the same as that for absorption $s \rightarrow r$, that is, $B_s^r = B_r^s$.

3. Thomson Scattering

The nonrelativistic limit of the Compton scattering of photons by free electrons is known as Thomson scattering. This limit applies when both the electron and photon momenta have magnitudes small compared with mc. If **p** and **p**′ denote the initial alhnal momenta of the electron and (**k**, λ) and (**k**′, λ′) denote the wave vectors and polarizations of the incident and scattered photons, then it follows from the laws of conservation of energy and momentum that $k' \approx k$ and hence that $p' \approx p$. Thus, the magnitudes of the momenta are effectively unaltered, although, in general, their directions change. In particular, for the limit considered, there is no shift in the frequency of the scattered photon.

The Feynman diagrams that give the leading contributions to Thomson scattering are shown in Fig. 8. Each diagram depicts the incident photon and initial electron

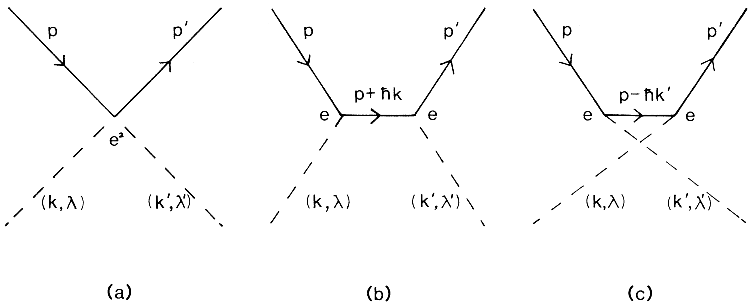


FIGURE 8 Feynman diagrams for Thomson scattering. Dotted lines represent photons and solid lines represent free electrons.

arriving from the left, and the scattered photon and recoil electron disappearing to the right. The contribution of Fig. 8a arises from the e^2 term in the interaction Hamiltonian; it may be said that here the incident photon is annihilated and the scattered photon simultaneously created. In Figs. 8b and 8c, on the other hand, the annihilation and creation are represented by two different one-photon vertices, each arising from the e term in the interaction Hamiltonian. These diagrams differ in the order in which the creation and annihilation take place. Overall momentum is conserved at every vertex.

It must be emphasized, however, that the contributions of Figs. 8a–8c cannot be physically separated, since only the initial and final states are observed. For this reason, the intermediate states are often referred to as virtual states. Indeed, it may be shown that the contributions of Figs. 8b and 8c effectively cancel, as their sum is of order $\hbar k/(mc)$ times that of Fig. 8a.

In scattering experiments the measured quantity is the cross section (having dimensions cm^2), defined as the number of scattered particles per unit time divided by the number of incident particles per unit area per unit time. For Thomson scattering, the differential cross section per unit solid angle Ω for scattering the photon with polarization λ' and direction within $d\Omega$ of $\hat{\mathbf{k}}'$ is given, from the contribution of Fig. 8a, by

$$\frac{d\sigma}{d\Omega} = r_0^2\left|\hat{\mathbf{e}}^{(\lambda)} \cdot \hat{\mathbf{e}}^{(\lambda')}\right|^2 = r_0^2 \cos^2 \Theta,$$

where (a) it is assumed that the polarization vectors are real, (b) Θ is the angle between the directions of polarization of the incoming and outgoing photons, and (c)

$$r_0 = e^2/4\pi mc^2 \approx 2.82 \times 10^{-13} \text{ cm}$$

and is the so-called classical electron radius. (This is the radius that an electron of uniformly distributed charge must have if its electrostatic energy is to equal its rest energy.) It should be noted in particular that, with the approximations indicated, the cross section is independent of frequency and vanishes if the incident and scattered polarization vectors are perpendicular.

If the incident photons are randomly polarized and the polarization of the scattered photons is not observed, then the cross section should be averaged over initial polarization indexes and summed over final polarization indexes. The resulting differential cross section per unit solid angle depends only on the scattering angle θ (where $\cos\theta = \hat{\mathbf{k}} \cdot \hat{\mathbf{k}}'$):

$$d\sigma/d\Omega = \tfrac{1}{2} r_0^2 (1 + \cos^2\theta).$$

The total unpolarized cross section, obtained by integrating this over all solid angles, is given by

$$\sigma_{\text{tot}} = (8\pi/3) r_0^2 \approx 6.65 \times 10^{-25} \text{ cm}^2.$$

4. Other Applications

The field of application of nonrelativistic quantum electrodynamics has expanded considerably in recent years due to the development of lasers and their use as spectroscopic tools for investigating a variety of physical and chemical systems. Lasers are sources of highly coherent and very intense beams of light. In the subject of quantum optics, the quantum statistical properties, such as the degree of coherence, of the light beam itself may be the object of investigation. For example, the quasi-classical states of the radiation field referred to earlier exhibit a higher degree of coherence than that of the occupation-number states, and these in turn are less chaotic than fields in thermal equilibrium, in which the distribution of photons follows Planck's radiation law.

The high intensity of the light beams that may be achieved by using laser sources can also give rise to nonlinear effects that are unobservable at lower intensities. A typical example of a nonlinear process is third-harmonic generation, that is, the absorption of three photons of frequency ν by an atom and the emission of a single photon of frequency 3ν (the third harmonic of the incident frequency). The rate for this process (after which the atom returns to its initial state and overall energy is consequently conserved) is proportional to the cube of the intensity of the incident beam. This should be contrasted with a linear process such as single-photon absorption for which the transition rate is proportional to the intensity itself, the factor of proportionality being Einstein's B coefficient.

III. RELATIVISTIC QUANTUM ELECTRODYNAMICS

A. Relativistic Theory

Relativistic quantum electrodynamics is formed by the union of the special theory of relativity, characterized by the speed of light, and quantum mechanics, characterized by Planck's constant. In discussions of the relativistic theory it is useful and customary to employ the natural system of units, in which speeds are measured as multiples of c and angular momenta are measured as multiples of $\hbar$. Since no natural length appears in the theory, lengths continue to be measured in centimeters. The expression for any quantity in (rationalized) natural units is obtained from the corresponding expression in (rationalized) cgs units simply by setting $c = 1$ and $\hbar = 1$. For example, the cgs expressions $\hbar\mathbf{k}$, mc^2, and $e^2/(4\pi\hbar c)$ for the momentum of a photon, the rest energy of an electron, and the fine-structure constant, respectively, become $\mathbf{k}$, m, and $e^2/(4\pi)$, respectively, in natural units. It is also easy to convert from natural units to cgs units, by inserting appropriate factors of $\hbar$ and c.

If quantum electrodynamics is to satisfy the principles of the special theory of relativity, its equations must be covariant under Lorentz transformations. Lorentz transformations relate the space–time coordinates x, y, z, and t of events as seen by observers using inertial frames of reference moving with uniform velocity relative to each other. (The coordinates x, y, z, and t are the components of a four-dimensional vector, to be denoted simply by x.) A covariant equation has the same form for two such observers. The fact that physical laws are expressible as covariant equations means that these laws are the same for all observers using inertial reference frames.

The state of a quantum-mechanical system (for example, the electromagnetic field *in vacuo*) is specified in relativistic theory on a three-dimensional spacelike hyperplane. In a given inertial frame, this consists of either all the points in three-dimensional space at a particular instant of time or all the events on a two-dimensional plane moving for all time perpendicularly to itself with constant speed greater than that of light. Two distinct events on a spacelike hyperplane cannot be connected by signals traveling with speed less than or equal to the speed of light, and so two measurements made in the vicinity of the corresponding space–time points will not interfere. This is known as microscopic causality. The whole of the four-dimensional space–time manifold is filled with a set of parallel spacelike hyperplanes, which may be labeled by an invariant timelike parameter τ, with τ ranging from $-\infty$ to ∞. The evolution of the system is described by specifying the state for each hyperplane τ and is determined dynamically, through Schrödinger's equation, on the interval $[\tau_1, \tau_2]$, if the state is specified at τ_1 and no measurements are made until τ_2.

B. Electrons and Positrons

The one-particle relativistic theory of the electron is based on the Dirac equation. This is a differential equation, with matrix coefficients, for a spinor wave function $\psi(x)$ having four components $\psi^\mu(x)$ $(\mu = 0, 1, 2, 3)$. The requirement that the Dirac equation be covariant determines the behavior of ψ under Lorentz transformations. Since the electron is now described by a four-component spinor rather than by a one-component scalar wave function, it has extra degrees-of-freedom, over and above those allowed by the Schrödinger theory. These correspond to the spin or intrinsic angular momentum of magnitude $\frac{1}{2}$ (in natural units). Hence the spin appears automatically in the Dirac theory of the electron and does not have to be added on in an *ad hoc* fashion, as it does in nonrelativistic quantum mechanics.

The difficulties of interpretation associated with the negative-energy solutions of the Dirac equation have already been mentioned. These difficulties disappear in the second-quantized version of the theory, in which electrons and positrons are treated on an equal footing and all have positive energy. Moreover, this version provides a calculus for processes involving annihilation and creation of electrons and positrons—in the high-energy regime, the number of these particles is no longer conserved.

The spinor ψ and its related adjoint spinor $\bar{\psi}$ are first expressed in terms of plane-wave solutions of the free-particle equation. These solutions correspond to particles with energy E, momentum $\mathbf{p}$, and rest mass e satisfying the relativistic energy–momentum relation

$$E^2 = |\mathbf{p}|^2 + m^2.$$

The coefficients in the expansions of ψ and $\bar{\psi}$ may then be interpreted as annihilation and creation operators for electrons and positrons in definite momentum and spin states. The spin states can be chosen to be helicity states of the electrons and positrons, that is, states in which the component of spin in the direction of motion is either $\frac{1}{2}$ (right-hand helicity) or $-\frac{1}{2}$ (left-hand helicity). It is only the component of spin in the direction of motion that is invariant under Lorentz transformations.

The algebra of the creation and annihilation operators for electrons and positrons differs from that of the creation and annihilation operators for photons. (Technically, it involves anticommutation instead of commutation relations.) The difference arises from the fact that, whereas photons are bosons, electrons and positrons are fermions and are subject to the exclusion principle. The only possible occupation numbers for electrons or positrons are therefore 0 or 1. This is in agreement with Pauli's spin-statistics theorem, since the spin of an electron or a positron is half an odd integer. It can be shown that quantizing the Dirac field by using commutation relations instead of anticommutation relations leads to a Hamiltonian operator with energy levels that are not bounded below and, hence, one for which no stable vacuum (ground) state exists. (Similarly, quantizing the Maxwell field, which has intrinsic spin 1, by using anticommutation relations instead of commutation relations leads to a breakdown of microscopic causality.)

C. Covariant Quantization of the Electromagnetic Field

The quantization of the electromagnetic field in the Coulomb gauge, though very useful for dealing with bound systems in nonrelativistic approximation, is not a manifestly covariant procedure. In the Coulomb-gauge formalism, only the transverse field is quantized, while the longitudinal field gives rise to an instantaneous interaction between the charges. The division of the field into transverse and longitudinal components, however, is not Lorentz covariant—these components do not transform separately on going from one inertial frame to another. So that the covariance of the theory can be exhibited, it is necessary to use a guage condition on the electromagnetic potentials that is itself covariant. The most convenient such condition is the so-called Lorentz condition. This condition also leads to certain difficulties which are, however, overcome in the formalism developed by Gupta and Bleuler.

The electromagnetic field is, in the first instance, quantized in a covariant way without reference to the Lorentz-gauge condition. In contrast to the noncovariant treatment, there are now, for each wave vector **k**, four types of photon, corresponding to timelike and longitudinal as well as two transverse polarization vectors, which are, in addition, four-dimensional rather than three-dimensional vectors. Moreover, the inner (or scalar) product of the infinite-dimensional vector space on which the photon creation and annihilation operators act is not positive definite; that is, there exist nonzero vectors in this space the square of whose length is zero or negative. (This is due to the metric of the space–time continuum, which distinguishes timelike from spacelike directions. Thus, the four-dimensional vector x is spacelike, lightlike, or timelike, relative to the origin, according to whether as $x^2 + y^2 + z^2 - t^2$ is positive, zero, or negative.) This constitutes a serious difficulty, since the quantum-mechanical statistical interpretation requires a positive-definite inner product. For the resolution of this problem, the use of the Lorentz-gauge condition, which has yet to be imposed, is of decisive importance.

It may be shown that neither the Lorentz condition nor Maxwell's equations are satisfied as operator equations in the covariant theory, because they are incompatible with the commutation relations. They are, however, satisfied as equations for expectation values (and hence are satisfied in the classical limit), provided a subsidiary condition is imposed on those state vectors that are to represent physically realizable states. The effect of the subsidiary condition is to make the timelike and longitudinal photons unobservable in real states of the system. These states have either no timelike or longitudinal photons at all or only certain allowed admixtures of them. Moreover, changing the allowed admixtures is merely equivalent to carrying out a gauge transformation that maintains the Lorentz condition. The allowed admixtures are always such that the contributions of timelike and longitudinal photons to, for example, the energy and momentum, cancel out, and only the contributions of the transverse, observable photons remain. Similarly, the statistical interpretation of the theory is consistent, when this is restricted to the calculation of probabilities for physically realizable states.

Despite the fact that timelike and longitudinal photons are unobservable in real states of the system, their presence is important and cannot be neglected in intermediate or virtual states. For example, the Coulomb interaction may be described in terms of the virtual exchange of timelike and longitudinal photons by charged particles. The appearance of these photons in the formalism is also required, of course, if the theory is to be manifestly Lorentz covariant.

D. Symmetries and Conservation Laws

The coupling between the quantized Maxwell and Dirac fields is represented by a Lorentz-invariant interaction Hamiltonian density $\mathscr{H}_{\mathrm{INT}}$ that links scalar and vector potentials to the charge and current densities. Here $\mathscr{H}_{\mathrm{INT}}$ is linear in the electromagnetic potentials and bilinear in the spinor fields ψ and $\bar{\psi}$ and is also of order e—there is no e^2 term as in nonrelativistic theory. The interaction Hamiltonian density may be derived from a Lagrangian density known as the minimal-coupling Lagrangian density.

It is interesting to note that certain continuous symmetries of the coupled systems are reflected in the structure of the complete Lagrangian density $\mathscr{L}$, which is Lorentz invariant and gauge invariant. According to a theorem of Noether, these symmetries must lead to conservation laws. For example, the invariance of $\mathscr{L}$ under time displacements implies the conservation of energy; its invariance under space displacements and rotations implies the conservation of linear and angular momentum, respectively; and its invariance under gauge transformations implies the conservation of charge.

The complete system also has three discrete symmetries. It is invariant under (a) charge conjugation C, that is, the interchange of particles and antiparticles (which affects only electrons and positrons, since the photon is its own antiparticle); (b) the parity operation P, that is, space inversion or the interchange of left and right; and (c) time reversal T. This invariance under C, P, and T is not shared by all the laws of nature. The nonconservation of parity in the weak interaction, which is responsible for the dynamics of beta emission, was suggested by Lee and Yang in 1956 and subsequently confirmed experimentally. That the combined transformation of charge conjugation and parity is also not a symmetry follows from the decay of the long-lived neutral K meson into two charged pions, a decay that is forbidden by CP conservation. Invariance under the conbined CPT transformation, established on very general assumptions (Lorentz covariance and locality), then implies that time reversal is also not a symmetry of the physical world. Hence, the separate conservation of C, P, and T is only an approximation which is, however,

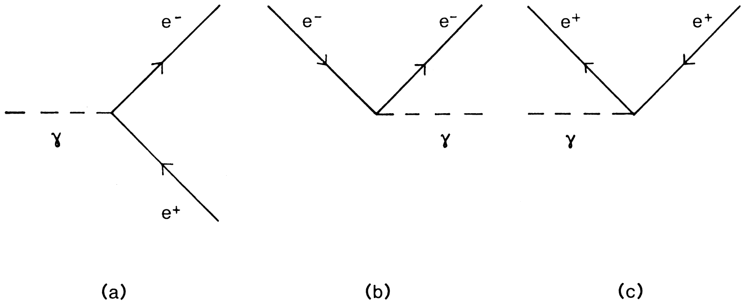


FIGURE 9 Virtual processes depicted by basic vertex diagrams (to be viewed from left to right). (a) Photon (γ) annihilation and electron–positron ($e^{-}e^{+}$) pair production. (b) Electron scattering and photon creation. (c) Photon annihilation and positron scattering. Note the convention for the sense of the arrows used on the fermion lines to distinguish electrons and positrons.

valid for phenomena that are adequately described by electrodynamics alone.

E. The *S* Matrix and Feynman Diagrams

The S matrix in quantum electrodynamics is used to calculate probability amplitudes for processes in which particles (electrons, positrons, or photons) that are initially free are allowed to interact and scatter. In the so-called interaction picture of the motion, the state vector Ψ for the complete system evolves under the influence of the interaction Hamiltonian $\mathcal{H}_{\text{INT}}$, alone, and the S operator (or scattering operator) maps the state vector on the hyperplane $\tau = -\infty$ (that is, long before the interaction takes place) onto the state vector on the hyperplane $\tau = \infty$ (that is, long after the interaction has ceased):

$$\Psi(\infty) = S\Psi(-\infty).$$

The S operator can be developed as a power series in the coupling constant e. With the help of a theorem due to Wick, the structure of the nth-order contribution, corresponding to the nth power of e in the expansion, may be systematically analyzed and represented by Feynman diagrams. It is usually convenient to use Feynman diagrams in energy–momentum space. These represent all possible virtual processes that can take place for given initial and final momentum and polarization or spin states of the particles. The Feynman rules enable expressions for the probability amplitude or S-matrix element S_{fi} for the process $i \to f$ to be written down directly from the diagrams. From this the cross section for the process may be calculated to a given order in e and compared with the experimentally obtained value.

The lowest order of perturbation theroy ($n = 1$) involves only the first power of the interaction Hamiltonia $\mathcal{H}_{\text{INT}}$, which is linear in the photon annihilation and creation operators and bilinear in the fermion (electron or positron) annihilation and creation operators. This gives rise to processes such as those depicted in the Feynman diagrams of Fig. 9. These diagrams are called basic vertex diagrams. There are in all eight such diagrams, corresponding to processes in which a photon is either annihilated or created and two fermions are annihilated or created or one is annihilated and the other created.

Every Feynman diagram is a combination of some or all of the eight basic vertex diagrams—an nth-order diagram contains n vertices. Energy and momentum (which together form a four-dimensional vector) are conserved at every vertex. (This is in contrast to the nonrelativistic theory, in which momentum but not energy is conserved in virtual processes.) However, the relativistic relation between energy and momentum need not be satisfied for virtual particles. Now this relation cannot be satisfied by all the particles participating in a basic vertex process, which must therefore be a virtual rather than a real process. For example, electron–positron annihilation with the production of a single photon is forbidden by energy–momentum conservation, even though it is allowed by charge conservation. Hence the basic vertex diagrams can appear only as parts of larger Feynman diagrams depicting processes for which overall energy and momentum are conserved and the relativistic energy–momentum relation is satisfied by the (real) particles in the initial and final states.

As an example of a real process, consider the Compton scattering of photons by electrons. This is allowed in the second order of perturbation theory, and the Feynman diagrams, each containing two vertices, are shown in Fig. 10. The corresponding polarized cross section for the laboratory reference system, in which the target electron is initially at rest, is given by the Klein-Nishina formula:

$$\frac{d\sigma}{d\Omega} = \frac{\alpha^2}{4m^2}\left(\frac{\nu'}{\nu}\right)^2\left[\frac{\nu}{\nu'} + \frac{\nu'}{\nu} + 4(\varepsilon \cdot \varepsilon)^2 - 2\right].$$

Here ν and ν' are the frequencies of the incident and scattered photons, respectively, and ε and ε' are their (four-dimensional) transverse polarization vectors, which in this

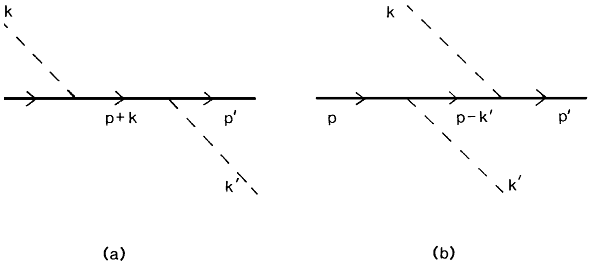


FIGURE 10 Feynman diagrams for Compton scattering. The lines are labeled by the four-dimensional energy–momentum vectors of the particles. Energy and momentum are conserved overall and at every vertex. Polarization and spin lables have been suppressed. Diagrams (a) and (b) differ in the order in which the incident photon is annihilated and the scattered photon created.

formula are assumed to be real (so that the photons are linearly polarized). In the low-energy limit ($\nu \ll m$ and $\nu' \approx \nu$), this reduces to the Thomson cross section derived from the nonrelativistic theory. (Note that $r_0 = \alpha/m$). The unpolarized cross section, obtained by averaging over initial and summing over final polarizations, is given by

$$\frac{d\sigma}{d\Omega} = \frac{\alpha^2}{2m^2}\left(\frac{\nu'}{\nu}\right)^2\left[\frac{\nu}{\nu'} + \frac{\nu'}{\nu} - \sin^2\theta\right],$$

where θ is the angle of scattering, as in Fig. 3. This reduces to the unpolarized Thomson cross section in the low-energy limit.

F. Radiative Corrections

The first approximation to the S-matrix element for a given process may be improved by adding contributions from higher-order perturbation theory. These contributions, known as radiative corrections, often, thought not always, involve integrals with ultraviolet divergences (that is, the integrals tend to infinity as the upper limits on the momenta of the virtual photons or fermions involved tend to infinity). For example, radiative corrections of second order in e (or of first order in α) relative to the lowest-order term are expected when one of the modifications shown in Fig. 11 is made in a Feynman diagram. Each of the integrals corresponding to the modified diagrams, however, has an ultraviolet divergence.

The divergence difficulties of relativistic quantum electrodynamics may be overcome by first regularizing the theory, that is, by altering it so that all the integrals converge. In the method of dimensional regularization, for example, this is achieved by replacing (in a well-defined sense) divergent four-dimensional expressions by convergent $(4-\varepsilon)$-dimensional expressions, where $\varepsilon > 0$. This may be described as reducing the dimensions of energy–momentum space from 4 to $4-\varepsilon$. The regularized theory is not equivalent to quantum electrodynamics, which is restored only in the limit as $\varepsilon \to 0$, in which limit the divergences reappear.

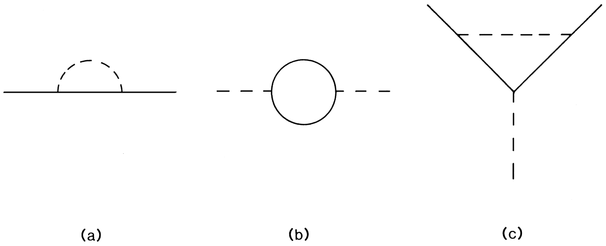


FIGURE 11 Modifications of a fermion line, a photon line, and a basic vertex part leading to second-order radiative corrections. (a) Fermion self-energy arising from emission and reabsorption of virtual photons. (b) Photon self-energy (or vacuum polarization) arising from virtual pair creation and annihilation. (c) Vertex modification arising from virtual photon exchange.

The mass and charge of the fermions are then renormalized; that is, the predictions of the regularized theory are expressed in terms of the observed mass and charge of the physical particles rather than the unobservable mass and charge of the bare particles. In this connection, certain relations between fermion self-energy and vertex-modification contributions (see Figs. 11a and 11c), known as Ward's identities, allow a great simplification to be made. In particular, they imply that charge renormalization arises solely from vacuum polarization effects (see Fig. 11b). It is important to note also that mass and charge renormalization would have to be carried out even if no divergences appeared in the formalism.

Finally, quantum electrondynamics is recovered by removing the regularization. If the method of dimensional regularization is used, this means taking the limit as $\varepsilon \to 0$. In this limit, infinities reappear in the relations between the observed and bare masses and charges. These relations, however, are not susceptible to experimental verification, as the bare masses and charges themselves are unobservable. Moreover, as $\varepsilon \to 0$, the physical predictions of the theory (for example, rediative corrections to scattering cross sections or electromagnetic shifts of energy levels) are finite in all orders of perturbation theory and are expressed in terms of the observed masses and charges. (For this reason quantum electrodynamics is said to be a renormalizable quantum field theory.) These predictions can therefore be tested against experimental results.

1. The Lamb Shift

The nonrelativistic Lamb shift for atomic hydrogen was first calculated by Bethe in 1947, following the experiments of Lamb and Retherford. Whereas Dirac's one-particle relativistic theory predicts that the $2S_{1/2}$ and $2P_{1/2}$ states of the hydrogen atom have the same energy, Lamb and Retherford showed that the $2S_{1/2}$ level is actually

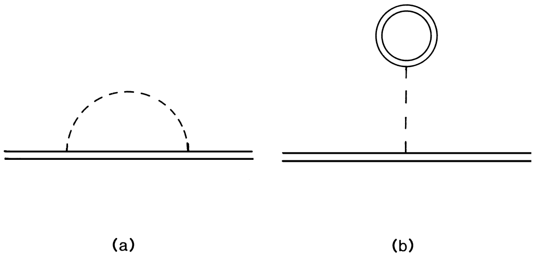


FIGURE 12 Feynman diagrams for second-order radiative corrections to atomic energy levels (Lamb shift). (a) Electron self-energy and (b) vacuum polarization.

higher and found a difference in energy corresponding to a frequency of about 1000 MHz. In Bethe's treatment, the effect was interpreted as a difference between the electron's self-energy when free (Fig. 11a) and when bound to the proton (Fig. 12a); a cutoff of order *mc* was used for the momenta of the virtual photons emitted in each case. The calculation gave no shift for the $2P_{1/2}$ level but did give an upward shift of about 1040 MHz for the $2S_{1/2}$ level and was therefore, in view of the nonrelativistic treatment and the approximations made, in good agreement with the observed value of the separation. This agreement has been enhanced by subsequent refinements in both theory and experiment.

The relativistic treatment of the Lamb shift, which is a bound-state problem, requires a more elaborate formalism (involving the so-called bound interaction picture) than the *S*-matrix theory outlined above. In addition to electron self-energy effects, there are vacuum polarization effects (see Fig. 12) and effects due to the finite mass and nonzero size of the nucleus. (The proton is not a pointlike object but has an effective radius of order 10^{-13} cm.) Vacuum polarization gives a downward shift of about 27 MHz to the $2S_{1/2}$ level. The two most precise directly measured values for the $2S_{1/2} - 2P_{1/2}$ level splitting in atomic hydrogen are 1057.845(9) MHz, obtained by Lundeen and Pipkin in 1981, and 1057.8514 (19) MHz, obtained by Palchikov, Sokolov, and Yakovlev in 1983. [The figures in parentheses represent uncertainties in the last digit(s) quoted.] In 1994, Hagley and Pipkin deduced the value 1057.839(12) indirectly, by measuring the $2S_{1/2} - 2P_{3/2}$ frequency interval and using the theoretical value for the $2P_{1/2} - 2P_{3/2}$ fine-structure splitting. A recent theoretical value for the Lamb shift, given by Pachucki *et al*. in 1997, is 1057.839(4)(4) MHz, where the first error is due to uncertainty in the value $0.862(12) \times 10^{-13}$ cm used for the proton radius and the second error stems from estimates of higher-order binding corrections. The uncertainties, both theoretical and experimental, in the Lamb-shift frequency are of the order of 10^4 Hz. This should be compared with a frequency of order 10^9 Hz for the Lamb shift itself and a frequency of order 10^{15} Hz for an optical transition.

2. The Anomalous Magnetic Moment of the Electron

The comparison of the measured and calculated values of the anomalous magnetic moment of the electron is regarded as an important test of quantum electrodynamics. The anomalous moment arises from small deviations of the electron's gyromagnetic ratio from 2, which is the value predicted by the Dirac theory. The gyromagnetic ratio g_{e^-} is defined through the relation between the intrinsic magnetic moment **M** of the electron and its spin angular momentum **S**, namely,

$$\mathbf{M} = -g_{e^-}\left(\frac{e}{2mc}\right)\mathbf{S}.$$

The directly measured quantity is not the gyromagnetic ration (or g-factor, as it is also called) itself but the electron anomaly a_{e^-}, which is the difference between this ratio and 2, all divided by 2. Thus,

$$a_{e^-} = \frac{g_{e^-} - 2}{2}.$$

The electron anomaly is, like the g-factor, a dimensionless constant. Its value is approximately one-tenth of one percent and has been both measured and calculated with great precision. The experimental and theoretical values of a_{e^-} agree to seven significant figures

$$a_{e^-} = 0.001159652.$$

The value of g_{e^-} is obtained from this by simple arithmetic:

$$g_{e^-} = 2.002319304.$$

In experiments carried out at the University of Washington in Seattle, the accuracy of the measurement has been greatly increased. In these experiments, electrons were held in a configuration of static electric and magnetic fields in a cavity with linear dimensions of order 1 cm. The arrangement is known as a Penning trap. Even single electrons can be held in it for weeks at a time. The electrons in a Penning trap have a discrete set of energy levels and are sometimes considered as part of an atom with a nucleus of macroscopic size, namely the experimental apparatus or, indeed, the earth on which it rests; the atom is called geonium.

The measured value of a_{e^-} was reported by Van Dyck, Schwinberg, and Dehmelt in 1987 as

$$a_{e^-}^{\text{exp}} = 0.0011596521884(43),$$

where again the figures in parentheses represent the probable uncertainty in the last digits. The electron anomaly, or the g-factor, is the most accurately known of all physical constants. Its measurement does not depend on a knowledge of either the values of other physical constants

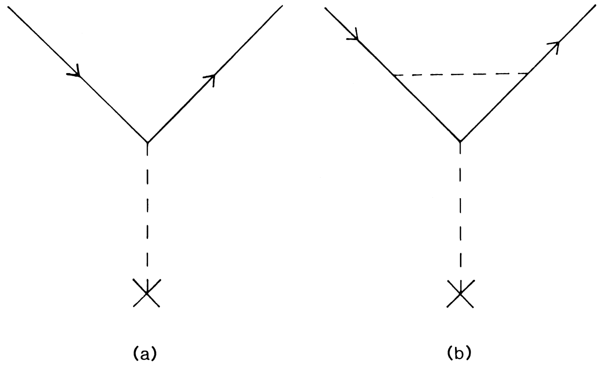


FIGURE 13 Feynman diagrams for electron scattering by an external field (denoted by X). (a) Zeroth-order contribution yielding a g-factor of 2 or an electron anomaly of 0. (b) Radiative correction of order α yielding a g-factor of $2+(\alpha/\pi)$ or an electron anomaly of $\alpha/2\pi$.

or the strength of the magnetic field involved in the experiment.

The theoretical value of a_{e^-} is obtained by considering the scattering of electrons by an external (prescribed) field. Feynman diagrams for the lowest-order contribution to this process and a radiative correction of order α are shown in Fig. 13. The change in the momentum of the scattered electron ish supplied by the external field. The lowest-order contribution to the electron anomaly is known exactly (its value $\alpha/2\pi$ was calculated by Schwinger in 1948), as is the contribution of order α^2. Further contributions of order α^3 and α^4 have also been calculated, partly analytically and partly numerically. (The contribution of order α^4 arises from 891 different Feynman diagrams.) The theoretical value given by

$$a_{e^-}^{\text{th}} = 0.001159652140(5.3)(4.1)(27.1)$$

was obtained by Kinoshita and Lindquist in 1990. Here the first and second errors are numerical and the third (and dominsant) error arises from uncertainties in the value of the fine-structure constant α.

The positron anomaly a_{e^+} was measured by Van Dyck, Schwinberg, and Dehmelt in 1987 by using the geonium experiment. They concluded that any difference between the ratio of the positron g-factor to the electron g-factor and unity must be less than 10^{-11}. This conclusion is strong evidence for the validity of the *CPT* theorem. Any departure of g_{e^+}/g_{e^-} from 1 would signal a breakdown of the combined charge conjugation, parity, and time-reversal transformation as a symmetry of nature.

G. Interaction of Photons and Leptons

Relativistic quantum electrodynamics may readily be extended to include the interaction of photons with certain other charged particles besides electrons and positrons. These are the muon (symbol μ^-) and the tauon (symbol τ^-) and their antiparticles μ^+ and τ^+. Muons and tauons have, within experimental accuracy, the same charge $(-e)$ and spin $(\frac{1}{2})$ as electrons, but different masses. While the rest energy (measured in electron volts) of the electron is about 0.511 MeV, that of the muon is about 105.659 MeV and that of the tauon is (with a possible error of about 3 MeV) about 1784 MeV. The fact that the electron, muon, and tauon seem to have identical characteristics (apart from mass) is known as $e-\mu-\tau$ universality. The muon and the tauon have lifetimes of order 10^{-6} sec and 10^{-13} sec, respectively. Electrons, muons, and tauons are all called leptons (as are neutrinos, which are, however, uncharged); they do not, in contrast to hadrons, experience the strong (nuclear) force. On the other hand, they do participate in the weak and gravitational interactions as well as in the electromagnetic interaction.

An example of a scattering process that involves more than one kind of lepton is muon pair production in electron–positron collisions. The Feynman diagram for the lowest-order contribution to this is shown in Fig. 14. An electron and a positron are annihilated and a virtual photon is created; this in turn is annihilated and a muon and an antimuon are created. For the process to occur, the electron and positron together must have at least the threshold energy equal to twice the rest energy of the muon (about 211 MeV). It should be noted that in Fig. 14 the lepton number, defined as the number of leptons minus the number of antileptons, is conserved at each vertex for both electrons and muons. This is true generally (and for tauons as well) and arises from the form of the interaction Hamiltonian. Each basic vertex involves only one type of lepton or antilepton. There are no vertices involving, for example, the annihilation of an electron and the simultaneous creation of a muon.

In the center-of-mass reference system (in which the electrons and positrons collide head-on with the same energy E), the threshold energy is reached when the particles

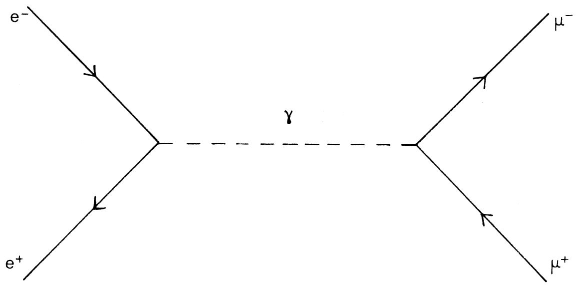


FIGURE 14 Muon pair production through electron–positron annihilation. The lepton number (in this case 0) is conserved at each vertex for each kind of lepton separately.

have been accelerated to speeds about 0.999988 times the speed of light. For energies much greater than the threshold energy (or speeds even closer to that of light), the unpolarized differential and total cross sections for muon pair production in the center-of-mass system reduce to

$$\frac{d\sigma}{d\Omega} = \frac{\alpha^2}{16E^2}(1 + \cos^2\theta)$$

and

$$\sigma_{\text{tot}} = \frac{\pi\alpha^2}{3E^2},$$

where θ is the angle between the incoming electron and the outgoing muon (or between the incoming positron and the outgoing antimuon). The second of these formulas has been verified to within a few percent in experiments using total center-of-mass energies of the order of 55 GeV. (At very high center-of-mass energies, weak-interaction effects must also be taken into account.)

The results of the high-energy experiments can be used to set bounds on possible deviations from exact quantum electrodynamics. The existence of heavy photons, for example, would modify the structure of the theory. (Thus, in the static limit, the Coulomb potential would no longer have a simple inverse-distance dependence). The total cross section for muon pair production would be altered according to

$$\sigma_{\text{tot}} \to \sigma_{\text{tot}}\left(1 \mp \frac{4E^2}{4E^2 - \Lambda_\pm^2}\right)^2,$$

where $\Lambda_\pm$ are cutoff parameters with the dimensions of energy. For consistency with the experimental results, $\Lambda_\pm$ must be at least of the order of 200 GeV. This corresponds to a test of the pointlike nature of photon–lepton interactions down to distances of the order of $1/\Lambda_\pm$, that is, less than 10^{-16} cm.

SEE ALSO THE FOLLOWING ARTICLES

Atomic Physics • Atomic Spectroscopy • Electromagnetics • Microoptics • Optical Diffraction • Quantum Optics • Quantum Theory • Relativity, Special • Unified Field Theories

BIBLIOGRAPHY

Cohen-Tannoudji, C., Dupont Roc, J., and Grynberg, G. (1989). “Photons and Atoms: An Introduction to Quantum Electrodynamics,” Wiley, New York.

Craig, D. P., and Thirunamachandran, T. (1998). “Molecular Quantum Electrodynamics: An Introduction to Radiation–Molecule Interactions,” Dover Mincola, N.Y.

Dowling, J. P., ed. (1997). “Electron Theory and Quantum Electrodynamics: 100 Years Later,” Plenum New York.

Feynman, R. P. (1985). “QED: The Strange Theory of Light and Matter,” Princeton Univ. Press, Princeton, N.J.

Greiner, W., and Reinhardt, J. (1994). “Quantum Electrodynamics,” 2nd ed. Springer-Verlag, Berlin.

Healy, W. P. (1982). “Non-Relativistic Quantum Electrodynamics,” Academic Press, London.

Kinoshita, T., ed. (1990). “Quantum Electrodynamics,” World Scientific, Singapore.

Milonni, P. W. (1994). “The Quantum Vacuum: An Introduction to Quantum Electrodynamics,” Academic Press, Boston.

Pike, E. R., and Sarkar, S. (1995). “The Quantum Theory of Radiation,” Clarendon, Oxford.

Weinberg, S. (1995). “The Quantum Theory of Fields,” Vol. 1, Cambridge Univ. Press, Cambridge.

Quantum Chemistry

Donald J. Kouri
University of Houston

GLOSSARY

Born-Oppenheimer approximation Approximate separation of nuclear and electronic variables in Schrödinger equation.

Configuration interaction Variational method using a linear combination of fixed Slater determinants.

Coupled cluster method Nonvariational method using virtual Hartree-Fock spin-orbitals to include configuration interaction effects.

Density functional theory Formally exact method expressing ground state energy in terms of total electron density.

Dividing surface "Surface of no return" that separates product arrangement from all other dynamical configurations.

Eulerian angles Three angles associated with orienting polyatomic systems in space.

Hartree-Fock method Variational method based on self-consistent field approximation, including antisymmetry effects.

Large Bra and Ket vectors Abstract states in quantum mechanics.

Møller-Plesset or many body perturbation theory Perturbation method with respect to a single Hartree-Fock determinant.

Multi-configuration Hartree-Fock method Variational method using a linear combination of Hartree-Fock determinants, in which both spin-orbitals and configuration coefficients are optimized.

Negative imaginary potential (NIP) Ad hoc potential introduced to absorb wave function in dynamical regions of no interest; generally located at dividing surfaces.

Pauli principle Requirement that wave functions be antisymmetric under exchange of identical, half-odd integral spin particles.

Perturbation method Computational method using solutions to a reference problem for a more complicated problem.

Quantum transition state theory Method using NIPs at dividing surfaces to calculate cummulative reaction probability in terms of a "strong interaction Green function."

Slater determinant Wave function expression guaranteeing antisymmetry.

Spin Relativistic effect giving rise to intrinsic angular momentum.

Time-independent wave packet theory Time-independent Schrödinger equation containing initial wave packet.

Transition state State of dynamical system at a dividing surface.

Variational method Computational method ensuring an upper bound to the system energy.

Wave function Projection of quantum state onto position eigenstate.

Wave packet Time-dependent wave function.

QUANTUM CHEMISTRY deals with the detailed understanding of atomic and molecular structure, properties, and dynamics based on the mathematical framework of quantum mechanics. This involves determining the quantum mechanical "state vector" of the system, and most commonly in chemistry, this is done in the so-called "coordinate representation," yielding the "wave function." The time evolution of the state vector or wave function is governed by the time-dependent Schrödinger equation. A typical strategy expresses the behavior and properties of a bulk system in terms of the smallest possible dynamical system displaying the underlying physical or chemical properties or processes. In the case of chemical reactions, this involves synthesizing the bulk reaction in terms of the smallest possible number of atomic and/or molecular species needed for the basic reaction stoichiometry. For spectroscopy, magnetic, electrical, etc. properties, it typically involves determining such properties for individual atoms and/or molecules.

Under the usual conditions, the most important interactions governing atomic and molecular structure, properties, and dynamics are electromagnetic, and time is treated nonrelativistically. There are, however, some important manifestations of special relativity, such as spin degrees of freedom and the Pauli exclusion principle. It should be clear that quantum chemistry has much in common with atomic and molecular physics and theoretical physics. Here, we focus on the areas of most intense research activity in chemistry.

I. CALCULATING ATOMIC AND MOLECULAR WAVE FUNCTIONS

A. Some Basics of Wave Mechanics

We assume that the reader has some familiarity with quantum mechanics, and here we summarize only a few of the most important aspects, primarily to establish notation. For any atomic or molecular system, there is an associated state vector, $|\chi\rangle$, whose time evolution is governed by the abstract Schrödinger equation,

$$\hat{H}|\chi\rangle = i\hbar\frac{\partial}{\partial t}|\chi\rangle. \tag{1}$$

We use the convenient Dirac "ket" notation, $|\chi\rangle$, for the state vector (along with the Dirac "bra" notation for the "dual space" vector, $\langle\chi|$, used to compute scalar products in quantum mechanics); $\hat{H}$ is the Hamiltonian operator (the generator of the state vector time evolution); and $\hbar$is Planck's constant h divided by 2π. In quantum chemistry, $\hat{H}$ usually consists of the sum of all contributions to the kinetic energy of the system, $\hat{K}$, and all contributions to the potential energy of the system, $\hat{V}$. The former takes account of energy due to motion, and the second takes account of energy due to position. Like all physical observables in quantum mechanics, $\hat{H}$ is a "self-adjoint" or Hermitian operator, ensuring real eigenvalues. Particularly useful are state vectors associated with specific values of a complete set of observables. Such sets of observables are composed of the maximum number of dynamical variables whose quantum operators commute (since commuting operators are not subject to a Heisenberg uncertainty condition). Two of the most useful sets of observables are the Cartesian positions of the particles in the system and their canonically conjugate momenta. As canonically conjugate variables, the position and corresponding momentum operators of a particle do not commute. Consequently, it is customary to use one or the other, and in quantum chemistry, the overwhelming importance of coulombic interactions makes the position variables most convenient. States of well-defined position are eigenvectors of the position operator, $\hat{x}$, so that for a particle confined to the x-axis,

$$\hat{x}\,|\,x\rangle = x\,|\,x\rangle, \tag{2}$$

where the eigenvalue x is the point on the axis where the particle is *exactly* located and $|x\rangle$ is the (improper) eigenvector for this state. The wave function, $\chi(x)$, is a measure of how likely it is that a particle in the state $|\chi\rangle$ will be found at the point x. This is called a probability amplitude and is computed as the scalar product of the state $|\chi\rangle$ with the state $|x\rangle$,

$$\chi(x) \equiv \langle x\,|\,\chi\rangle. \tag{3}$$

If one also represents the Hamiltonian operator in terms of its effect on $\chi(x)$, we expect the kinetic energy operator in the coordinate representation to result in changes in where the particle is likely to be found. The most general form is given by

$$\langle x|\hat{K}|\chi\rangle = \int_{-\infty}^{\infty} dx'\langle x|\hat{K}|x'\rangle\langle x'\,|\,\chi\rangle. \tag{4}$$

The coordinate representation of the kinetic energy operator is

$$\langle x|\hat{K}|x'\rangle = -\frac{\hbar^2}{2m}\delta(x-x')\frac{\partial^2}{\partial x^2}, \tag{5}$$

where $\delta(x - x')$ is the Dirac delta function. For the potential energy, we expect $\hat{V}$ to be a function of the particle position operator, $\hat{x}$, so in the coordinate representation it is diagonal, with eigenvalues given by the classical potential energy function, $V(x)$. Then the coordinate representation of the time-dependent Schrödinger equation [in one dimension (1D)] is

$$\left[-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}+V(x)\right]\chi(x,t)=i\hbar\frac{\partial}{\partial t}\chi(x,t). \qquad (6)$$

For typical chemical systems, Eq. (6) applies to each Cartesian coordinate for each electron and nucleus in the system, and there are also potential energy terms for the interparticle coulombic interactions. The general form is

$$\begin{aligned} e^2\Bigg\{\Bigg(&\sum_{l>l'}\sum_{l'=1}^{L-1}\frac{1}{|\vec{r}_l-\vec{r}'_{l'}|}+\sum_{a>a'}\sum_{a'=1}^{N-1}\frac{Z_aZ_{a'}}{|\vec{R}_a-\vec{R}_{a'}|}\\ &-\sum_{a=1}^{N}\sum_{l=1}^{L}\frac{Z_a}{|\vec{R}_a-\vec{r}_l|}\Bigg)\Bigg\}\chi(\{\vec{R}_a,\vec{r}_l\},t)\\ &\times\left\{-\frac{\hbar^2}{2}\left(\sum_{a=1}^{N}\frac{1}{m_a}\nabla_a^2+\frac{1}{m_e}\sum_{l=1}^{L}\nabla_l^2\right)\right\}\chi(\{\vec{R}_a,\vec{r}_l\},t)\\ &=i\hbar\frac{\partial\chi}{\partial t}. \qquad (7)\end{aligned}$$

Here, (a, a') label nuclei, (l, l') label electrons, $\{\vec{R}_a, \vec{r}_l\}$ denotes the set of all vectors from a common origin to the nuclei and electrons, ∇_a^2 is the three-dimensional Laplacian for nucleus a (having charge Z_a), and ∇_l^2 is the three-dimensional Laplacian for electron l. The enormous difficulty associated with solving Eq. (7) basically stems from the coulombic repulsions and attractions which prevent a separation of variables. In recent years, quantum chemists have become increasingly interested in solving the time-dependent Schrödinger equation directly. However, it remains true that the overwhelming majority of computations have focussed on taking advantage of the fact that (in the absence of external, time-varying perturbations) $\hat{H}$ is independent of any explicit time dependence. This makes possible solution by separation of variables,

$$\chi(\{\vec{R}_a,\vec{r}_l\},t)=\Psi(\{\vec{R}_a,\vec{r}_l\})\Phi(t), \qquad (8)$$

where

$$i\hbar\frac{\partial}{\partial t}\Phi=E\Phi, \qquad (9)$$

$$H\Psi=E\Psi. \qquad (10)$$

Note that we shall always use Dirac notation when writing abstract state vectors and denote abstract operators with a caret. Wave functions and coordinate representation operators will be indicated by ordinary Greek and Roman letters, and we shall suppress the explicit coordinate dependence, except when it is required for understanding. The separation constant, E, is interpreted as the total energy of the system, and Eq. (10) is the "workhorse" of quantum chemistry. However, it is perhaps useful to point out that while the above is totally general for systems in which the preparation of the system is irrelevant (e.g., systems not involving sources), it is not for other problems. For example, in reactive scattering experiments one may desire to determine highly resolved information (e.g., state-to-state cross sections at specific energies), and one *must* know the detailed initial conditions in order to make a comparison between experiment and theory. Although seldom discussed, the special nature of Eq. (10) can be seen by deriving it in a more general manner.

We do this by noting that time and energy are conjugate variables in the sense of Fourier analysis. A more general way to derive a time-independent wave function equation is to Fourier transform Eq. (7), written more compactly as

$$H\chi=i\hbar\frac{\partial}{\partial t}\chi. \qquad (11)$$

If we specify an experiment lasting from t_i to t_f, a natural definition of a time-independent wave function is

$$\Psi\equiv\frac{1}{\sqrt{2\pi}}\int_{t_i}^{t_f}dte^{iEt/\hbar}\chi. \qquad (12)$$

The equation determining Ψ is obtained, in general, by applying H to Ψ, using Eq. (11), and integrating by parts to yield

$$(E-H)\Psi=-\frac{i\hbar}{\sqrt{2\pi}}\left[e^{iEt_f/\hbar}\chi(t_f)-e^{iEt_i/\hbar}\chi(t_i)\right]. \qquad (13)$$

Now t_f is the end time of the experiment, and for scattering, the products, whatever they are, will have passed through the detector and exited the apparatus. Therefore, $\chi(t_f)$ is essentially zero inside the apparatus. However, to obtain Eq. (10), one must also assume that $\chi(t_i)$ is also zero everywhere within the apparatus, but this contradicts the fact that one *must* measure what it is in order to analyze the experimental data. Therefore, the time-independent Schrödinger equation that corresponds to experiment must be

$$(E-H)\Psi=\frac{i\hbar}{\sqrt{2\pi}}e^{iEt_i/\hbar}\chi(t_i). \qquad (14)$$

It is convenient to define t_i to be zero. Thus, the time-independent Schrödinger equation retains a memory of the experimental details. This should not be surprising since it is well known in quantum statistical mechanics that the time-independent von Neumann equation contains the initial density matrix. Of course, the ultimate goal is to determine dynamical information characteristic of the chemical species involved, independent of the experimental details. This means that the ultimate quantities calculated will have had all reference to the experiment removed. Individual definite energy, state-to-state cross sections from Eqs. (7), (10), or (14) will be the same. However, there can be computational advantages to solving equations containing specific experimental conditions.

B. General Aspects of Nuclear and Electronic Dynamics: Born-Oppenheimer Approximation

Except for the hydrogen atom and its isotopic variants, there are no closed form analytical solutions to the Schrödinger equation for atoms and molecules. The simplest molecules are H_2^+ and H_2 (and isotopic variants), and they provide convenient examples for sketching the difficulties encountered. For H_2^+, Eq. (10) is

$$\left\{-\frac{\hbar^2}{2m_p}\left[\nabla^2_{\vec{R}_1} + \nabla^2_{\vec{R}_2}\right] - \frac{\hbar^2}{2m_e}\nabla^2_{\vec{r}_1} + \frac{e^2}{|\vec{R}_1 - \vec{R}_2|} - \frac{e^2}{|\vec{R}_1 - \vec{r}_1|} - \frac{e^2}{|\vec{R}_2 - \vec{r}_1|}\right\}\Psi = E\Psi, \qquad (15)$$

and the coulombic attraction terms prevent separation of the $\vec{r}_1$ dependence. However, the mass of the electron is about 1837 times smaller than that of a proton, so at any given kinetic energy, electrons travel some 43 times more slowly. This suggests one can neglect changes in the nuclear wave function on the time scale for electron dynamics. Mathematically, this is manifested by the approximate separation of variables in Eq. (15), and one should be able to solve the electron dynamics at fixed $|\vec{R}_2 - \vec{R}_1|$, with the wave function written in a product form. In fact, one may do this in a much more careful way. The potential depends only on the three distances separating the two protons and the electron from each nucleus. It is independent of the three coordinates of the molecular center of mass and the three (Eulerian) angles orienting the three particle triangle. Therefore, one rigorously separates out these six coordinates. The remaining dynamics is then characterized by the total angular momentum quantum number J, the component of angular momentum quantum number M with respect to a z-axis fixed at the system center of mass with arbitrary orientation, and the component of angular momentum quantum number Λ with respect to a "body-fixed" z-axis that rotates to maintain a definite orientation with respect to the three particle triangle. The wave function becomes a vector of functions, Ψ^J_Λ, satisfying equations of the form

$$\sum_{\Lambda'=-J}^{J} H^J_{\Lambda\Lambda'}\Psi^J_{\Lambda'} = \left(E - \frac{e^2}{R}\right)\Psi^J_\Lambda, \qquad (16)$$

where the Hamiltonian matrix contains the radial kinetic energy and centrifugal energy for the relative nuclear motion, the centrifugal energy for rotation of the electron about the interproton axis, and terms that are nondiagonal in the Λ index (referred to as "coriolis coupling," describing the tumbling in space of the three particle triangle). It is common to neglect the coriolis coupling so that one solves a single uncoupled equation for each Ψ^J_Λ:

$$H^J_{\Lambda\Lambda}\Psi^J_\Lambda = \left(E - \frac{e^2}{R}\right)\Psi^J_\Lambda. \qquad (17)$$

This equation is still nonseparable, and the Born-Oppenheimer approximation consists of assuming a product solution

$$\Psi^J_\Lambda(R, \xi, \eta) = \zeta^J_\Lambda(R)\phi_\Lambda(R, \xi, \eta) \qquad (18)$$

and neglecting $\frac{\partial\phi_\Lambda}{\partial R}$ and $\frac{\partial^2\phi_\Lambda}{\partial R^2}$ terms. Here, ξ, η denote the remaining electron coordinates. Then one obtains an equation of the form

$$\phi_\Lambda H^J_{a\Lambda}(R)\zeta^J_\Lambda(R) + \zeta^J_\Lambda(R)H_{e\Lambda}(\xi, \eta, R)\phi_\Lambda = \left(E - \frac{e^2}{R}\right)\zeta^J_\Lambda(R)\phi_\Lambda(R, \xi, \eta), \qquad (19)$$

where $H^J_{a\Lambda}(R)$ is a radial kinetic energy operator for the relative dynamics of the two protons (nuclei); it also contains the centrifugal potential (if any) associated with the internuclear rotation. We solve for the eigenvalues and eigenfunctions of the electronic Hamiltonian $H_{e\Lambda}(\xi, \eta, R)$:

$$H_{e\Lambda}\phi_{n\Lambda} = \epsilon_{n\Lambda}(R)\phi_{n\Lambda}. \qquad (20)$$

Note that the electronic eigenenergies depend on R through the electron-proton attraction terms of the potential operator in $H_{e,\Lambda}$. Its dependence on Λ takes account of the electronic rotation about the internuclear vector. Also, the quantum number Λ enters $H_{e\Lambda}$ as Λ^2, so each energy level $\epsilon_{n\Lambda}$ is twofold degenerate. The potential energy for nuclear motion associated with the Born-Oppenheimer state $\zeta^J_\Lambda(R)\phi_{n\Lambda}(R, \xi, \eta)$ is

$$U_{n\Lambda}(R) = \epsilon_{n\Lambda}(R) + \frac{e^2}{R}. \qquad (21)$$

The electronic energy $\epsilon_{n\Lambda}(R)$ and the associated wave function $\phi_{n\Lambda}(R, \xi, \eta)$ can be thought of as "adiabatic" energies and states. They essentially assume that the electron adjusts adiabatically to the nuclear motion. Inclusion of the nuclear derivatives, $\frac{\partial}{\partial R}\phi_\Lambda$ and $\frac{\partial^2}{\partial R^2}\phi_\Lambda$, leads to states that are viewed as "diabatic." More rigorous treatments exist in which these nuclear kinetic energy terms are eliminated by a mathematical transformation defining the adiabatic and diabatic states.

The last step in the procedure (at the Born-Oppenheimer level) is to solve the Schrödinger equation for the nuclear vibrational motion,

$$\left[H^J_{n\Lambda}(R) + U_{n\Lambda}(R)\right]\zeta^J_{n\Lambda\nu}(R) = \epsilon_{J\nu}\zeta^J_{n\Lambda\nu}(R), \qquad (22)$$

and for scattering dynamics,

$$\left[H^J_{n\Lambda}(R) + U_{n\Lambda}(R)\right]\zeta^{J+}_{n\Lambda}(E \mid R) = E\zeta^{J+}_{n\Lambda}(E \mid R). \qquad (23)$$

Corrections to take account of the nonadiabatic effects of nuclear-electronic coupling require inclusion not only of the $\frac{\partial}{\partial R}\phi_{n\Lambda}$, $\frac{\partial^2}{\partial R^2}\phi_{n\Lambda}$ terms, but also the coriolis coupling in the Λ quantum number, as well as inclusion of more than just a single product form for $\Psi^J_\Lambda(R, \xi, \eta)$. Thus, Eq. (18) is replaced by

$$\Psi^J_\Lambda(R, \xi, \eta) = \sum_n \zeta^J_{n\Lambda}(R)\phi_{n\Lambda}(R, \xi, \eta), \qquad (24)$$

which is substituted into Eq. (16) to generate coupled equations for the $\zeta^J_{n\Lambda}(R)$. The $\phi_{n\Lambda}$ are members of the complete set of eigenfunctions of the Hamiltonian $H_{e\Lambda}$, in Eq. (20). The study of electronic nonadiabatic coupling is currently of great interest in quantum chemistry.

Finally, before going on, we display in Fig. 1 some examples of the potential energy functions, $U_{n\Lambda}(R)$, governing the relative nuclear dynamics. The electronic states can be understood in part by using their behavior in two extreme limits. The "united atom" limit ($R \to 0$) results in a total nuclear charge $(Z_1 + Z_2)e$, so for H_2^+, it gives a charge of $+2e$. The lowest H_2^+ electronic state then correlates with the ground state of He^+. When the nuclear repulsion, $\frac{e^2}{R}$ is added to obtain $U(R)$, it results in a coulombic singularity at $R = 0$. At large R, H_2^+ yields a neutral H-atom and H^+, so the electronic energy tends to the ground state of the hydrogen atom. Since $\frac{e^2}{R} \to 0$ in this limit, we see that $U \to -0.5$ a.u. Similar considerations apply to excited states. For H_2^+, the first excited united atom state is the $2p_z$ level of He^+, and the large R limit is again a single ground state H-atom plus a proton. Of course, the protons are *identical* and one cannot distinguish to which one the electron is bound. This is reflected in the fact that the ground state, $\phi_{n=1\Lambda=0}(R \to \infty)$, wave function has the form

$$\phi_{1,0}(R \to \infty) = \frac{1}{\sqrt{2}}\left[\phi_{1s}(A) + \phi_{1s}(B)\right], \qquad (25)$$

where $\phi_{1s}(i)$ is the $1s$-hydrogen orbital centered on nucleus $i = A$ or B. The positive sign reflects the fact that the lowest electronic wave function should have the longest possible deBroglie wavelength and must be normalizable. This latter condition implies that $\phi_{1,0}$ vanishes for the electron infinitely far from both nuclei, and the former condition implies that it have no other "nodal surfaces." The first excited electronic state must likewise have the nodal surface at infinity (for normalizability), but it must also have an additional nodal surface (to give a shorter deBroglie wavelength). This can be ensured by writing

$$\phi_{2,0}(R \to \infty) = \frac{1}{\sqrt{2}}\left[\phi_{1s}(A) - \phi_{1s}(B)\right], \qquad (26)$$

since at any value of R this is zero when the electron is anywhere in the plane bisecting the internuclear distance. In the united atom limit, this plane contains the united nucleus, and $\phi_{1s}(A)$ provides the positive lobe and $-\phi_{1s}(B)$ provides the negative lobe of the $2p_z$ wave function. Similar ideas apply in general.

In the case of the H_2 molecule, there is an additional electron, which introduces additional complexity. This occurs because of an effect of the special theory of relativity, namely, electron spin and the requirements of Fermi-Dirac statistics and the Pauli exclusion principle. The relativistic requirement that time must be treated on an equal footing with the spatial coordinates was shown by Dirac to lead to an additional matrix structure for relativistically covariant quantum mechanics. It was also found that even in the nonrelativistic limit, this matrix structure (though somewhat simplified) did not go away. In this limit, Pauli showed that the matrix operators associated with the remaining relativistic effects obey the commutation relations of a quantum mechanical angular momentum operator, with eigenvalue for the square of the spin being $s(s+1)\hbar^2$, $s = \frac{1}{2}$, and $s_z = \pm\frac{1}{2}\hbar$ As a result, it was found necessary to multiply the usual nonrelativistic, single electron solutions of the Schrödinger equation by either of two spin states, $|\alpha\rangle$ and $|\beta\rangle$. By convention, the α-state corresponds to $s_z = +\frac{\hbar}{2}$ and the β-state corresponds to $s_z = -\frac{\hbar}{2}$. The energy (Hamiltonian) is independent of the spin angular momentum if one neglects electromagnetic interactions associated with the electron's magnetic moments arising due to its orbital and spin angular momentum. (Classically, any current or charge circulation produces a magnetic field.) In an analogous fashion, nuclei that are fermions also possess intrinsic (spin) angular momentum. For example, protons and neutrons are fermions, having half-integral spin. Consequently, associated with the nuclear part of an atomic or molecular wave function one must include nuclear spin states.

The second complication arising if a system contains more than one of a given type of fermion (e.g., two protons in H_2^+; two protons and two electrons in H_2; three electrons in Li; etc.) is that the wave function (or state vector) must change sign if any two identical particles are exchanged. So long as the system only contains two of any identical fermion particles, the effects of antisymmetry can be accounted for by forming a product of an appropriately symmetrized and antisymmetrized spatial and spin wave functions.

In the case of H_2, one approach is to use single electron molecular orbitals patterned after the electronic states of H_2^+. Then a symmetric spatial wave function consists of each electron occupying a $\phi_{1,0}$-like state, with the total spatial wave function being a product $\phi_{1,0}(1)\phi_{1,0}(2)$, with the index in parentheses labeling electron 1 or 2. Obviously, this wave function is symmetric under exchange of the electrons. The spin part of the wave function must be antisymmetric,

$$\chi_{s=0,s_z=0}(1,2) = \frac{1}{\sqrt{2}}[\alpha(1)\beta(2) - \alpha(2)\beta(1)], \qquad (27)$$

or in the usual determinantal form,

$$\chi_{s=0,s_z=0} = \frac{1}{\sqrt{2}}|\alpha(1)\alpha(2)\beta(1)\beta(2)|. \qquad (28)$$

In general, interchanging the pair of electrons interchanges the columns of the determinant, which ensures a change of sign. This state corresponds to one electron with $s_z = \frac{\hbar}{2}$ and one with $s_z = -\frac{\hbar}{2}$. The total S_z is zero, as is also the magnitude $\hat{S}^2$. The complete molecular orbital description of ground state H_2 is then

$$\phi_{\Lambda=0,S=0} = \phi_{1,0}(1)\phi_{1,0}(2)\chi_{S=0,S_z=0}(1,2). \qquad (29)$$

This is a simple example of the "linear combination of atomic-molecular orbital" (LCAO-MO) description of the electronic structure of H_2. It is of interest to multiply out the spatial part of this LCAO-MO wave function:

$$\begin{aligned}\phi_{1,0}(1)\phi_{1,0}(2) = \frac{1}{2}[&\phi_{1s}(A,1)\phi_{1s}(A,2) + \phi_{1s}(B,1)\phi_{1s} \\ &\times (B,2) + \phi_{1s}(A,1)\phi_{1s}(B,2) \\ &+ \phi_{1s}(B,1)\phi_{1s}(A,2)]. \qquad (30)\end{aligned}$$

This illustrates a very important aspect of the LCAO-MO approach, since the terms $\phi_{1s}(i,1)\phi_{1s}(i,2),\ i = A,\ B$ represent *both electrons bound to the same nuleus,* A or B. These are called "ionic terms" since, e.g., the state $\phi_{1s}(A,1)\phi_{1s}(A,2)$ implies that B has lost an electron which is now bound to A.

Another approach is to assign an electron to a spin-orbital so that, e.g., having electron 1 on nucleus A with spin α is represented by $\phi_{1s}(A,1)\alpha(1)$. Similarly, electron 2 on nucleus B with spin β is described by $\phi_{1s}(B,2)\beta(2)$. We can construct an antisymmetrized total electron wave function as the "Slater determinant"

$$\begin{aligned}\tilde{\phi}_{\Lambda=0,S=0} = \frac{1}{\sqrt{2}}|&\phi_{1s}(A,1)\alpha(1)\phi_{1s}(A,2)\alpha(2) \\ &\times \phi_{1s}(B,1)\beta(1)\phi_{1s}(B,2)\beta(2)|. \qquad (31)\end{aligned}$$

Expanding this yields

$$\begin{aligned}\tilde{\phi}_{\Lambda=0,S=0} = \frac{1}{\sqrt{2}}[&\phi_{1s}(A,1)\alpha(1)\phi_{1s}(B,2)\beta(2) \\ &- \phi_{1s}(B,1)\beta(1)\phi_{1s}(A,2)\alpha(2)]. \qquad (32)\end{aligned}$$

Clearly, this wave function contains *no* ionic terms, but rather assigns one electron to nucleus A and one electron to nucleus B. This is an example of a "valence bond" electronic wave function. In general, both ionic and covalent type terms contribute to the accurate description of electronic states in molecules. In a general LCAO-MO treatment, one would use a sum of a number of atomic electronic configurations to form molecular orbitals, and by doing so, arbitrarily accurate results could (in principle) be obtained. Similarly, one could take a sum of many distinct valence bond wave functions so that many electronic configurations contribute. Again (in principle), arbitrarily accurate results can be obtained. Both LCAO-MO and valence bond approaches are used by quantum chemists.

When more than two identical Fermions are present in an atom or molecule (e.g., Li or LiH), one no longer can construct a properly antisymmetrized wave function as a product of a symmetrized total spatial and an antisymmetrized total spin wave function. This essentially results from the fact that there are only two possible single electron spin states, α and β. If one tries to form a spin determinant for three (or more) electrons, one *always* finds two rows of the determinant are equal and it vanishes identically. The consequence is that one must form determinants using one-electron products of a spin and spatial function. Any particular spatial function can be multiplied by either an α or β spin state so that, at most, each spatial orbital can be occupied by no more than two electrons. The spatial function can be either an LCAO-MO or a valence bond function.

One last important aspect of molecular wave functions is the inclusion of the effect of identical nuclei. Nuclei with half-odd integral spin are Fermions, and the *total* wave function (nuclear plus electronic) must be antisymmetric under exchange of any two identical nuclei. If they have integer spin, they are Bosons, and the total wave function must be symmetric under exchange of two such nuclei. In the case of H_2, we note that exchanging protons A and B affects the nuclear wave function through the polar and azimuthal angles of the internuclear vector. The symmetry is determined by whether J is even or odd, with odd (even) J yielding a nuclear rotational wave function that is odd (even) under nuclear exchange. However, it is also clear that the LCAO-MO $\phi_{1,0}$ [see Eq. (25)] is even if A and B are exchanged, but $\phi_{2,0}$ [see Eq. (26)] will change sign under this interchange. If both electrons in H_2 occupy either the $\phi_{1,0}$ or $\phi_{2,0}$ spatial orbitals, then the total electronic wave function is even under proton exchange. If one occupies $\phi_{1,0}$ and the other occupies $\phi_{2,0}$, then the electronic wave function is odd under the exchange. Then the even (odd) J-rotational states times the even (odd) total electronic wave function requires an antisymmetric nuclear spin wave function, while the odd (even) J-rotational states times even (odd) total electronic states require a symmetric nuclear spin state. We note that the odd nuclear spin state has the form

$$\chi_{S=0,S_z=0}(A,B) = \frac{1}{\sqrt{2}}\left[\alpha(A)\beta(B) - \beta(A)\alpha(B)\right], \quad (33)$$

and the symmetric spin states are

$$\chi_{S=1,S_z=1} = \alpha(A)\alpha(B), \quad (34)$$

$$\chi_{S=1,S_z=-1} = \beta(A)\beta(B), \quad (35)$$

$$\chi_{S=1,S_z=0} = \frac{1}{\sqrt{2}}[\alpha(A)\beta(B) + \alpha(B)\beta(A)]. \quad (36)$$

Clearly, there are three states possible for $S = 1(-1 \leq S_z \leq 1)$ so it is the *triplet* spin state. The proper enumeration of the possible spin states with even- and odd-J nuclear rotational states is crucial in quantum statistical mechanics.

If there are three or more nuclei in the molecule and some or all are identical, appropriate generalizations of the above are necessary. In addition, there can be additional symmetry operations (for the nuclei in their equilibrium positions), and in general, it is computationally advantageous to take this into account both in the molecular electronic wave function and in the nuclear wave function. The use of group theory plays a major role in atomic and molecular spectroscopy.

The final topic in this discussion is the nature of the nuclear vibrational Schrödinger equation [Eq. (22)]. If the potential $U_{n\Lambda}(R)$ possesses a well sufficiently deep to support bound eigenstates, then one seeks to compute the $\zeta^{J}_{n\Lambda\nu}$ and $\epsilon_{J\nu}$. One commonly used technique is to expand $U_{n\Lambda}(R)$ in a Taylor series about the potential minimum at R_e:

$$U_{n\Lambda}(R) = U_{n\Lambda}(R_e) + \frac{1}{2}\frac{d^2}{dR^2}U_{n\Lambda}|_{R_e}(R - R_e)^2, \quad (37)$$

since $\frac{d}{dR}U_{n\Lambda}|_{R_e} \equiv 0$. In addition, the centrifugal potential is approximated by its value at $R = R_e$. This gives the energy $\epsilon_{J\nu}$ as a sum of the harmonic oscillator energy $\hbar\omega_e(\nu + \frac{1}{2})$ and the rigid rotor energy $J(J+1)\hbar^2/2\mu R_e^2$, where μ is the reduced mass of the two nuclei and ω_e is the usual harmonic oscillator angular frequency. A more accurate description can be obtained by including more terms in the Taylor expansion of both $U_{n\Lambda}$ and the centrifugal potential. This leads to a power series in the quantum numbers ν and J, with cross terms that arise from vibrational-rotational coupling, higher powers of $(\omega_e + \frac{1}{2})$ due to anharmonic effects, etc.

Again, analogous generalizations exist for systems with more than two nuclei. For systems with large numbers of nuclei, it is *not* convenient nor particularly useful computationally to separate the center-of-mass motion or the three Eulerian angles. This is essentially because of the enormous number of possible relative coordinates and rotating frames for such systems.

C. Computational Tools for Bound States

The two primary tools for computing energy levels and wave functions in quantum chemistry are the variational and perturbation theoretic methods. A key aspect of the wave functions for bound systems is the fact that since the probability of observing system components at infinite separations from one another must be zero, the wave function must vanish on the system boundaries. Further, because the probability density for observing the system in any configuration equals the modulus of the wave function squared, and the total probability of finding the system somewhere must equal 1, bound state wave functions must be "quadratically integrable" or $\mathcal{L}^2$-functions. In general, the value of the observable total energy of a system in state Ψ is given by

$$E = \int d\tau \Psi^* H \Psi \Big/ \int d\tau |\Psi|^2, \quad (38)$$

where $d\tau$ is the differential volume element and the integral is over all space. The Rayleigh-Ritz variational principle states that the functional variation of E with respect to either Ψ or Ψ^* vanishes, and that the extremum solution to the resulting Euler-Lagrange equations is a minimum. Rearranging Eq. (38), we calculate

$$\delta E \int d\tau |\Psi|^2 + \int d\tau \delta\Psi^*(E - H)\Psi$$
$$+ \int d\tau \Psi^*(E - H)\delta\Psi \equiv 0, \quad (39)$$

and we impose the condition $\delta E \equiv 0$. Since complex Ψ and Ψ^* are linearly independent, so also are their variations and consequently the coefficients of both $\delta\Psi$ and $\delta\Psi^*$ must separately vanish. This yields the usual Schrödinger equation, $(E - H)\Psi = 0$, or its complex conjugate, and establishes that a complete minimization of E yields the exact energy. Alternatively, we can expand Ψ in Eq. (38) using the complete set of eigenfunctions of H, $\{\phi_n\}$, satisfying

$$H\phi_n = E_n\phi_n, \quad (40)$$

so that

$$\Psi = \sum_n c_n \phi_n. \tag{41}$$

Then subtracting the lowest (ground state) energy, E_0, from both sides of Eq. (38), and substituting Eq. (41), we obtain

$$E - E_0 = \sum_n |c_n|^2 (E_n - E_0) \Big/ \sum_{n'} |c_{n'}|^2. \tag{42}$$

The right-hand side of the above equation is greater than or equal to zero, so we conclude that

$$E \geq E_0, \tag{43}$$

which proves that Eq. (38) yields an upper bound to the ground state energy of the system. In practice, one typically expands Ψ in some appropriate basis functions (satisfying the same boundary conditions as the exact Ψ):

$$\Psi = \sum_j a_j \psi_j. \tag{44}$$

Then E is computed, along with $\frac{\partial E}{\partial a_j^*}$ (noting that since H is self-adjoint, it is not necessary to also examine $\frac{\partial E}{\partial a_j}$), yielding

$$E \sum_{j'} S_{jj'} a_{j'} = \sum_{j'} H_{jj'} a_{j'}, \tag{45}$$

with

$$H_{jj'} = \int d\tau \psi_j^* H \psi_{j'} \equiv \langle \psi_j | H | \psi_{j'} \rangle, \tag{46}$$

and

$$S_{jj'} = \int d\tau \psi_j^* \psi_{j'} \equiv \langle \psi_j \mid \psi_{j'} \rangle. \tag{47}$$

The matrix $S_{jj'}$ is called the "overlap" matrix; if the basis functions are orthonormal, $S_{jj'} = \delta_{jj'}$, where the Kronnecker delta $\delta_{jj'}$ is zero unless $j = j'$. In matrix notation, Eq. (45) is referred to as the generalized eigenvalue problem of linear algebra:

$$E\mathbf{S} \cdot \mathbf{a} = \mathbf{H} \cdot \mathbf{a}. \tag{48}$$

In addition, normalization of Ψ implies that

$$\sum_j a_j^* a_j = 1. \tag{49}$$

Major research activities in quantum chemistry center around developing efficient ways to compute the $H_{jj'}$ and to solve for the E's and $\mathbf{a}$'s.

In perturbation theory, the strategy is to use or construct a solvable quantum problem which is as close as possible to the unsolvable system of actual interest. If H_0 is the "reference system" Hamiltonian and H is the true system Hamiltonian, then the perturbation, λV, is defined as

$$\lambda V = H - H_0, \tag{50}$$

where λ is an arbitrary parameter such that when $\lambda = 1$, one has the true system. Let the complete set of eigenstates of H_0 be denoted by $\{\psi_j^0; j = 0, \ldots\}$, such that

$$H_0 \psi_j^0 = E_j^0 \psi_j^0. \tag{51}$$

It is assumed that the true eigenstates, satisfying

$$(H_0 + \lambda V)\psi_k = E_k \psi_k, \tag{52}$$

can be expanded in a power series in λ, as can E_k also:

$$\psi_k = \sum_{n=0}^{\infty} \lambda^n \psi_k^n, \tag{53}$$

$$E_k = \sum_{n=0}^{\infty} \lambda^n E_k^n. \tag{54}$$

We substitute Eqs. (53)–(54) into Eq. (52) and require the coefficients of each separate power of λ to vanish identically:

$$H_0 \psi_k^n + V \psi_k^{n-1} = \sum_{l=0}^{n} E_k^l \psi_k^{n-l}. \tag{55}$$

Thus,

$$H_0 \psi_k^0 = E_k^0 \psi_k^0, \tag{56}$$

$$H_0 \psi_k^1 + V \psi_k^0 = E_k^1 \psi_k^0 + E_k^0 \psi_k^1, \tag{57}$$

$$H_0 \psi_k^2 + V \psi_k^1 = E_k^2 \psi_k^0 + E_k^1 \psi_k^1 + E_k^0 \psi_k^2, \tag{58}$$

etc. The first equation shows that the zeroth order solution is given by the reference problem. The second equation leads to

$$\psi_k^1 = -\frac{1}{E_k^0 - H_0} \left(E_k^1 - V \right) \psi_k^0. \tag{59}$$

Inserting the resolution of the identity in the unperturbed basis, this becomes

$$\psi_k^1 = -\sum_{j=0}^{\infty} \frac{\psi_j^0}{E_k^0 - E_j^0} \left(E_k^1 \delta_{jk} - V_{jk} \right), \tag{60}$$

where we *must* require that

$$E_k^1 = V_{kk} \tag{61}$$

in order for ψ_k^1 to exist. Then, Eq. (60) becomes

$$\psi_k^1 = \sum_{j \neq k} \frac{\psi_k^0}{E_k^0 - E_j^0} V_{jk}, \tag{62}$$

and the first order correction to the energy is given by Eq. (61). If there are degeneracies, then Eq. (62) will *still* contain singularities for all j such that $E_j^0 = E\,_k^0$. When this occurs, the procedure is modified by replacing the degenerate ψ_j^0 by linear combinations:

$$\tilde{\psi}_k^0 = \sum_{j=\text{degenerate with } k} b_j^k \psi_j^0 \tag{63}$$

and choosing the $\{b_j^k\}$ so that the submatrix $\langle \tilde{\psi}_j^0 | V | \tilde{\psi}_{j'}^0 \rangle$ is *diagonal*. This (along with normalization of the new reference states) determines a new modified subset of degenerate, unperturbed states which are uncoupled by the perturbation. If the new states are labeled by $\tilde{\psi}_k^0$, then $\langle \tilde{\psi}_k^0 | V | \tilde{\psi}_{k'}^0 \rangle = 0, k \neq k'$, and the first order correction to the energy for the degenerate states is

$$E_k^1 \equiv \tilde{V}_{kk}, \tag{64}$$

and

$$\psi_k^1 = -\sum_{j \neq \{k\}} \frac{\psi_j^0}{E_k^0 - E_j^0} V_{jk}. \tag{65}$$

Here, $\{k\}$ denotes the modified degenerate states.

The second order correction is obtained by solving Eq. (59),

$$\psi_k^2 = -\frac{1}{E_k^0 - H_0}\left(E_k^0 \psi_k^0 - V \psi_k^1\right). \tag{66}$$

Again expressing the unperturbed Green function in the unperturbed basis (including the superposed degenerate states $\tilde{\psi}_k^0$), we obtain

$$\psi_k^2 = -\sum_j \frac{\psi_j^0}{E_k^0 - E_j^0}\left(E_k^2 \delta_{jk} - \langle \psi_j^0 | V | \psi_k^1 \rangle\right). \tag{67}$$

We require that

$$E_k^2 = \langle \psi_k^0 | V | \psi_k^1 \rangle \tag{68}$$

and

$$\psi_k^2 = \sum_{j \neq k} \frac{\psi_j^0}{E_k^0 - E_j^0} \langle \psi_j^0 | V | \psi_k^0 \rangle, \tag{69}$$

and we understand that no degeneracy terms occur where E_j^0 can equal E_k^0. The second order correction to the energy can be computed knowing the first order correction to the wave function. In fact, knowing ψ_k^1 enables one to determine the energy up to third order, E_k^3, according to

$$E_k^3 = \langle \psi_k^1 | V | \psi_k^1 \rangle - E_k^1 \langle \psi_k^1 | \psi_k^1 \rangle. \tag{70}$$

In general, knowing the wave function through order n determines the energy up through order $2n+1$.

II. ELECTRONIC STRUCTURE AND PROPERTIES OF ATOMS AND MOLECULES

A. Variational Methods

The most widely used approaches to calculating energies and wave functions for electrons in atoms and molecules have their roots in attempts to approximate the relevant Schrödinger equation (within the Born-Oppenheimer approximation) in terms of some sort of "independent electron" model. The crudest is simply to neglect all electron-electron repulsions and construct a product wave function using hydrogenic-like orbitals (with proper nuclear charge). Each orbital is assigned to two electrons until all electrons are accounted for. This includes the effect of the Pauli exclusion principle and leads to the simplest "aufbau principle" for building up atomic and molecular structure. More realistic results are obtained by the Hartree self-consistent field approach. Again, the wave function is taken to be a single product of one-electron orbitals, each associated with at most two electrons, but the orbitals are now considered unknowns. They are determined by successively projecting the action of the Hamiltonian on the product wave function onto all but one of the various orbitals, to obtain a set of effective one-electron orbital equations of the form

$$\left(-\frac{\hbar^2}{2m_e}\nabla_i^2 - \frac{Ze^2}{r_i} + V_i^{eff}\right)\phi_i = \epsilon_i \phi_i, i = 1, \ldots, Z, \tag{71}$$

where we illustrate the expression for a neutral atom with nuclear charge Ze. Here, V_i^{eff} is given by

$$V_i^{eff} = \sum_{j \neq i} \int d\vec{r}_j \phi_j^*(\vec{r}_j) \frac{e^2}{|\vec{r}_j - \vec{r}_i|} \phi_j(\vec{r}_j). \tag{72}$$

We see that Eq. (71) is a set of *nonlinear*, integro-differential equations that must be solved iteratively. This is done by guessing initial functions for the $\phi_i, i = 1, \ldots, Z$ that appear in the effective potentials. Then the resulting equations are uncoupled, linear partial differential equations for the ϵ_i and ϕ_i. Once they are solved, new V_i^{eff} are calculated and used to generate updated equations, Eq. (71), and the procedure is repeated until the ϵ_i, ϕ_i do not change to within a given tolerance. The approximate total energy is computed as the expectation value of the full Hamiltonian using the final product wave function determined by the self-consistent procedure.

A more sophisticated approach is the multiconfiguration self-consistent field Hartree-Fock (MCSCF-HF) approximation, which expresses the electronic wave function as a sum of "Slater determinants," constructed from one-electron spin-orbitals. This yields a properly antisymmetrized wave function,

$$\Psi = \sum_{l} C_l \Phi_l. \tag{73}$$

The individual "configuration" Slater determinants, Φ_l, are constructed from spin-orbitals, ϕ_j, given by

$$\phi_j = |\alpha\rangle \sum_{\lambda} C_{\lambda j} \chi_\lambda \quad \text{or} \quad |\beta\rangle \sum_{\lambda} C_{\lambda j} \chi_\lambda, \tag{74}$$

where χ_λ is a spatial basis function. The χ_λ generally fall into two classes:

- Slater-type orbitals (STOs)
$$\chi_\lambda = N_{nlm\zeta} Y_{lm}(\theta, \phi) r^{n-1} \exp(-\zeta r), \tag{75}$$
where $N_{nlm\zeta}$ is the normalization constant and ζ determines the radial "size" of the STO (physically, it is associated with an "effective nuclear charge").
- Cartesian Gaussian-type orbitals (GTOs)
$$\chi_\lambda = N_{abc\alpha} x^a y^b z^c \exp(-\alpha r^2), \tag{76}$$
where $N_{abc\alpha}$ is the normalization constant and α determines the radial size of the orbital. Note, e.g., that $(a, b, c) = (1, 0, 0),\ (0, 1, 0)$, or $(0, 0, 1)$ are $p_x,\ p_y,\ p_z$ orbitals; $(a, b, c) = (2, 0, 0),\ (0, 2, 0),\ (0, 0, 2),$ $(1, 1, 0),\ (0, 1, 1),\ (1, 0, 1)$ are sufficient to describe five d-orbitals and an s-orbital [since the sum of $(2, 0, 0),\ (0, 2, 0)$, and $(0, 0, 2)$ yields the function $r^2 \exp(-\alpha r^2)$].

In early work in quantum chemistry, STOs were the most commonly used basis, but as digital computers have become more powerful, with increased memory, the GTOs have dominated computations. The STOs have the attractive feature of behaving properly at the nucleus, but the matrix elements of the Hamiltonian between STO spin-orbitals are computationally costly. The GTOs permit these matrix elements to be evaluated routinely (even in polyatomic systems), but they do *not* behave correctly at the nucleus. However, it is possible to replace the χ_λ in Eq. (76) by a sum of Gaussians with fixed coefficients and different α-values, determined so as to mimic the cusp behavior at the nucleus. These are referred to as "contracted GTOs" (CGTOs). It is also sometimes efficient to use CGTOs that have been determined variationally from atomic structure calculations. These are useful for treating "core" electrons that retain a significant atomic character within the molecule.

In order to understand the level of sophistication of a given calculation, it is necessary to understand some basic terminology. A *minimal basis* is one in which the number of STOs or GTOs is equal to the number of core and valence orbitals in each atom. A *double zeta basis* is one in which there are twice as many STOs or GTOs as there are core and valence atomic orbitals. This gives greater flexibility to the LCAO-MOs that ultimately appear in each of the Slater determinants (configurations). Typically, one of the pair has a smaller exponent (α or ζ) and one has a larger exponent. A *triple zeta basis* has three times as many STOs or GTOs as the number of core and valence atomic orbitals. Even larger basis sets are possible. One such basis is denoted by $(10s, 6p/5s, 4p)$, indicating that 10 s-type primitive GTOs were contracted to yield 5 distinct s-type CGTOs and 6 primitive p-type GTOs were contracted to provide 4 distinct p-type CGTOs. Another Gaussian-type basis is denoted STO-3G, where three Gaussians each are used in a least squares fit of STOs, which have been optimized to describe various atomic electronic states. Yet others are denoted as 4-31G and 5-31G bases. In this case, the core orbital space is treated with CGTOs containing four or five Gaussians. The valence orbital space is treated at a double zeta level, with the first CGTO constructed from three primitive GTOs and the second containing one primitive GTO.

The framework of the Hartree-Fock approach is common to many of the most widely used electronic structure computational methods. Here we summarize some of these and indicate how they are related. In the MCSCF version, the value of $E = \langle\Psi|H|\Psi\rangle/\langle\Psi \mid \Psi\rangle$ is determined variationally with respect to the CI coefficients, C_l, in Eq. (73) and *simultaneously* with respect to the $C_{\lambda j}$ coefficients in the individual spin-orbitals. In addition, the $C_{\lambda j}$ are constrained to satisfy the ortho-normalization relations

$$\sum_{\lambda,\lambda'} C^*_{\lambda j} \langle\chi_\lambda \mid \chi_{\lambda'}\rangle C_{\lambda' j'} = \delta_{jj'}. \tag{77}$$

The fact that the exact H contains at most two-electron interactions implies that its matrix elements, while quadratic in the C_l, are quartic in the spin-orbital coefficients, $C_{\lambda j}$. This makes a full MCSCF calculation very costly and limits the size basis set that is practical. The CI method involves first determining the LCAO-MO coefficients in a *single determinant* SCF (sometimes called an "unrestricted Hartree-Fock" or UHF approximation) or small MCSCF calculation. Then the $C_{\lambda j}$ are fixed and a CI calculation is carried out in which *only* the C_l are varied.

B. Nonvariational Methods

There are several nonvariational methods that make use (usually) of the results of an UHF calculation. One method is called M-Plesset perturbation theory (MPPT) or, equivalently, many body perturbation theory (MBPT). In the UHF calculation, there are no C_l coefficients and the $C_{\lambda j}$ and ϵ_j are obtained by solving the so-called Roothan matrix Hartree-Fock equations,

$$\sum_{\lambda} F^{j}_{\lambda'\lambda} C_{\lambda j} = \epsilon_j \sum_{\lambda} S_{\lambda'\lambda} C_{\lambda j}, \tag{78}$$

where

$$\begin{aligned} F^{j}_{\lambda'\lambda} = \langle \chi_{\lambda'} | h | \chi_{\lambda} \rangle + \sum_{\delta\kappa} \big[& \gamma_{\delta\kappa} \langle \chi_{\lambda'} \chi_{\delta} | g | \chi_{\lambda} \chi_{\kappa} \rangle \\ & - \gamma^{exch}_{\delta\kappa} \langle \chi_{\lambda'} \chi_{\delta} | g | \chi_{\kappa} \chi_{\lambda} \rangle \big]. \end{aligned} \tag{79}$$

Here, h is the one-electron part of the full Hamiltonian, g is an electron-electron repulsion potential energy, and

$$\gamma_{\delta\kappa} = \sum_{i}{}' C_{\delta i} C_{\kappa i}, \tag{80}$$

$$\gamma^{exch}_{\delta\kappa} = \sum_{i}{}'' C_{\delta i} C_{\kappa i}, \tag{81}$$

where the single primed sum is over all occupied spin-orbitals and the doubly primed sum is over all occupied spin-orbitals having the same spin as spin-orbital j. Since the operator depends on the unknown $\{C_{\lambda j}\}$, the equations again must be solved iteratively. One specifies an initial guess for the occupied $\{C_{\lambda j}\}$; computes the Roothan Hartree-Fock matrix, $F^{j}_{\lambda'\lambda}$, $j = 1, \ldots, Z$; and then solves the resulting linearized eigenvalue equations [Eq. (78)] for the $\{\epsilon_j\}$ and the new $\{C_{\lambda j}\}$. Then new $\{F^{j}_{\lambda'\lambda}\}$ are computed and the procedure is repeated until the results converge. We note that the dimensionality of the matrix $F^{j}_{\lambda'\lambda}$ is $M \times M$, where M equals the total number of atomic basis orbitals used in the LCAO-MO expansion. It therefore has M eigenvalues ϵ_j and associated eigenvectors, $C_{\lambda j,m}$, $m = 1, \ldots, M$. In general, $1 \le j \le Z$, and M will be larger than Z. Consequently, only the lowest energy of each spin-orbital set, $\{\phi_{j,m}\}$, $1 \le j \le Z$, will be occupied. The unoccupied ones are referred to as *virtual spin-orbitals*.

The unperturbed Hamiltonian for the MPPT/MBPT method is taken to be the sum of the F^j matrices:

$$H_0 \equiv \sum_{j=1}^{Z} F^j, \tag{82}$$

and the perturbation is the difference between the exact H and H_0 in Eq. (82).

The other nonvariational approach using the UHF as its starting point is the coupled cluster (CC) method. In this approach, CI-like effects are included in a different fashion. One writes the total wave function as

$$\Psi = \exp(T)\Phi, \tag{83}$$

where Φ is usually an UHF Slater determinant and the operator T generates the so-called single, double, etc. "excitations." The singles are determinantal wave functions in which one of the occupied spin-orbitals is replaced by a virtual spin-orbital, doubles have two occupied spin-orbitals replaced by virtual spin-orbitals, etc. The form of T is

$$T = \sum_{i,m} t_{im} m^{+} i + \sum_{i,j,m,n} t_{ijmn} m^{+} n^{+} i j + \cdots, \tag{84}$$

where $i, j, \ldots$ remove occupied spin-orbitals and m^{+}, $n^{+}, \ldots$ create occupied virtual spin-orbitals $\phi_m, \phi_n, \ldots$. The $t_{im}, t_{ijmn}, \ldots$ play the role of the usual CI coefficients, C_I. However, they are *not* determined variationally. Rather, one uses the fact that $\Phi = \exp(-T)\Psi$ to write the Schrödinger equation as

$$\exp(-T) H \exp(T) \Phi \; = \; E \Phi. \tag{85}$$

This is projected onto the various single, double, etc. Slater determinants, $\{\Phi^{m}_{i}\}$, $\{\Phi^{mn}_{ij}\}, \ldots$ to yield, e.g.,

$$\left\langle \Phi^{m}_{i} | \exp(-T) H \exp(T) | \Phi \right\rangle = 0, \tag{86}$$

$$\left\langle \Phi^{mn}_{ij} | \exp(-T) H \exp(T) | \Phi \right\rangle = 0, \tag{87}$$

etc. The zeroes on the right-hand sides of Eqs. (86) and (87) result from the fact that all of the single, double, etc. Slater determinants are automatically orthogonal to the UHF state, $|\Phi\rangle$. This is a consequence of the occupied and virtual spin-orbitals for any j being eigenvectors of a Hermitian matrix.

It is useful to comment here on the strengths and weaknesses of variational and nonvariational approaches. In fact, the two approaches are complementary, in that variational methods provide rigorous upper bounds for the total energy, but these energies are *not* "size extensive." That is, a calculation on two CH_3 radicals at very large separation will not, in general, give an energy that is twice that of a single CH_3 radical. The MPPT/MBPT method *is* size extensive, but does *not* guarrantee that the total energy lies below the approximate energy. This suggests that the method of choice depends on the objective of a given application; e.g., extended systems are better treated nonvariationally. For following certain chemical reactions, electron rearrangement requires the inclusion of more than one configuration. In that case, the crucial role of the UHF determinant in MPPT/MBPT suggests that one would be better off using MCSCF or CI.

C. Density Functional Theory

Another major approach to atomic and molecular electronic structure is density functional theory (DFT). Closely related are the Thomas-Fermi and X—alpha (X_α) methods; since they have been supplanted by DFT, we shall concentrate on it. The fundamental basis of DFT is the fact that one can prove that the *ground state* energy, E_0, of a Z-electron system (atom or molecule) is *exactly* determined by a functional, $E_0(\rho)$, of the total electron density,

$\rho(\vec{r})$, of the system. Note that this does *not* provide the explicit dependence of E_0 on ρ, but rather guarantees that if one can determine ρ, this, in principle, determines the exact ground state energy. The resulting equations to be solved are

$$\left[-\frac{\hbar^2}{2m_e}\nabla^2 - \sum_n \frac{Z_n e^2}{|\vec{R}_n - \vec{r}|} + e^2 \int d\vec{r}' \frac{\rho(\vec{r}')}{|\vec{r} - \vec{r}'|} + U_{X_\alpha}(\vec{r})\right]\phi_i = \epsilon_i \phi_i, \tag{88}$$

where $U_{X_\alpha}(\vec{r})$ is an effective one-particle potential that accounts *exactly* (in principle) for all electron-electron correlation and exchange effects, and the integral term involving ρ is the coulombic interaction at the point $\vec{r}$ due to the total electron density. Note that U_{X_α} also must remove the "self-interaction" of the electron occupying orbital ϕ_i with its own contribution to $\rho(\vec{r})$. The DFT equations, like the Hartree approximation equations, must be solved iteratively, since it is easily verified that

$$\rho(\vec{r}) = \sum_j n_j |\phi_j|^2, \tag{89}$$

where n_j is the number of electrons occupying orbital ϕ_j. Thus, one begins with an initial guess of the ϕ_j and computes ρ. If one also has an estimate of U_{X_α}, then Eq. (88) is a linear eigenvalue problem for calculating a revised estimate of the ϕ_j and its single particle energy, ϵ_i. Once Eq. (88) is solved, a new $\rho(\vec{r})$ is calculated from Eq. (89) and the procedure is repeated until one reaches self-consistency. The total energy of the system is computed as

$$E = \sum_j n_j \left\langle \phi_j \middle| \frac{\hbar^2}{2m_e}\nabla^2 \middle| \phi_j \right\rangle - \int d\vec{r} \sum_n \frac{\rho(\vec{r}) Z_n e^2}{|\vec{R}_n - \vec{r}|} + \frac{e^2}{2}\int d\vec{r} \int d\vec{r}' \frac{\rho(\vec{r})\rho(\vec{r}')}{|\vec{r} - \vec{r}'|} + \frac{1}{2}\int d\vec{r} \frac{1}{2} U_{X_\alpha}(\vec{r}). \tag{90}$$

The greatest difficulty in the theory is obtaining an accurate approximation for U_{X_α}, and this is a very active area of research. A popular choice is the so-called "local density approximation" (LDA), for which

$$U_{X_\alpha} = -9\alpha/[2(3\rho(\vec{r})/8\pi)^{1/3}]. \tag{91}$$

Theoretical work indicates that $\alpha = 2/3$ is the optimum value to use.

All of the above discussion has been couched in purely theoretical terms. When all of the relevant matrix elements of the Hamiltonian are computed from first principles, the calculation is referred to as "ab initio." However, a great deal of quantum chemical applications use various ways to estimate the Hamiltonian matrix elements. These can range from ignoring the effects of nonorthogonality of the atomic orbitals centered on nonbonded atoms to approximating matrix elements using experimental data. The resulting approaches are referred to as semi-empirical methods, and there exist highly developed computer codes and data banks for carrying out such computations. Details can be found in some of the bibliographic references at the end of this article.

D. Other Properties

Finally, we comment on the calculation of properties other than the electronic energy. These are usually expressed in terms of the response of the system to some appropriate external perturbation, λV. These properties are typically evaluated by doing a Taylor expansion of the total energy of the perturbed system about $\lambda = 0$,

$$E = E(\lambda = 0) + \lambda \left(\frac{dE}{d\lambda}\right)_{\lambda=0} + \cdots \tag{92}$$

where

$$E(\lambda = 0) = \langle \Psi(\lambda = 0)|H(\lambda = 0)|\Psi(\lambda = 0)\rangle, \tag{93}$$

$$\begin{aligned}\left(\frac{dE}{d\lambda}\right)_{\lambda=0} &= \langle \Psi(\lambda = 0)|V|\Psi(\lambda = 0)\rangle \\ &+ 2\sum_l \left(\frac{\partial C_l}{\partial \lambda}\right)_{\lambda=0} \left\langle \frac{\partial \Psi}{\partial C_l}|H(\lambda = 0)|\Psi(\lambda = 0)\right\rangle \\ &+ 2\sum_{j,\mu} \left(\frac{\partial C_{\mu j}}{\partial \lambda}\right)_{\lambda=0} \left\langle \frac{\partial \Psi}{\partial C_{\mu j}}|H(\lambda = 0)|\Psi(\lambda = 0)\right\rangle \\ &+ 2\sum_{\mu} \left(\frac{\partial \chi_\mu}{\partial \lambda}\right)_{\lambda=0} \left\langle \frac{\partial \Psi}{\partial \chi_\mu}|H(\lambda = 0)|\Psi(\lambda = 0)\right\rangle,\end{aligned} \tag{94}$$

etc. The value of $(\frac{dE}{d\lambda})_{\lambda=0}$ depends not only on the nature of the perturbation, but also on the method used to obtain $\Psi(\lambda = 0)$. Thus, for the MCSCF method, derivatives of Ψ with respect to the C_l and $C_{\mu j}$ vanish identically since the MCSCF functional is stationary with respect to such variations. If the atomic orbitals, χ_μ, do not depend explicitly on the external field, then $(\frac{\partial \chi_\mu}{\partial \lambda})_{\lambda=0} \equiv 0$, and the first order effect is due solely to the average of V. In, e.g., the CI method, $\frac{\partial \Psi}{\partial C_l} \equiv 0$, but $\frac{\partial \Psi}{\partial C_{\mu j}} \neq 0$ since the $C_{\mu j}$ are not varied to minimize the CI energy functional.

III. QUANTUM CHEMICAL DYNAMICS

A. Early Transition State Theory

Over the past 30 or so years, the field of rigorous quantum chemical dynamics has been created. Prior to this, the dominant method used to calculate rates was the transition state approach. The role of quantum mechanics was mainly to determine the Born-Oppenheimer potential surface for nuclear dynamics and the eigenenergies of various bound degrees of freedom of the full system at the transition state (for use in calculating the transition state partition function). The definition of the transition state is most easily given for systems in which the collision process has to pass through a so-called "bottleneck" (e.g., assumed in many cases to be the saddlepoint in the Born-Oppenheimer nuclear potential surface, separating reactants and products). Generally, it is assumed that in all collisions in which the system reaches the transition state configuration, having positive kinetic energy (positive velocity in the direction of the product region), the reactive probability equals 1. The transition state is sometimes couched in terms of a "dividing surface" normal to which a positive velocity inexorably leads to reaction. An additional quantum modification is to introduce an additional "transition factor" which can account for effects such as tunneling (for collision energies below the saddlepoint barrier height, which otherwise have a zero reaction probability) as well as corrections to the assumption that all positive kinetic energy trajectories reaching the transition state cross to products with unit probability.

B. Rigorous State-to-State Quantum Reactive Dynamics

Although such transition state-based methods continue to play a major role in ab initio predictions of reaction rates, there now exist several rigorous quantum treatments of atom-diatom and diatom-diatom reactive collisions, yielding detailed quantum state resolved probabilities, cross sections, and rates. Most of these employ a rotating, center-of-mass coordinate frame (rigorously separating out the three coordinates of the center of mass and the three Euler angles characterizing the overall rotation of the three or four atom system). In reactive scattering, the major complication is the result of the fact that the natural coordinates for describing the atomic and molecular species in each arrangement are different. This is easily seen by considering an atom-diatom collision system like $D + HF$. Depending on the collision energy, there can be as many as four arrangements:

- $D + HF$
- $H + DF$
- $F + HD$
- $F + H + D$

In the first, the three natural coordinates remaining after separating off the center-of-mass and Euler angles of rotation are the internuclear HF distance, r_1; the distance of D from the HF center of mass, R_1; and the angle, γ_1, between these two distance coordinates. Analogous but distinct sets of three natural coordinates apply to the second and third rearrangements. For the last arrangement (termed the "breakup" arrangement), the natural coordinates are called hyperspherical coordinates and typically consist of a hyperradius ρ (whose square equals the sum of the squares of the two distance coordinates for any of the other three arrangements, i) and two angles. One of these can be the same as the γ_i in any of the other coordinate sets, and the second angle is typically defined as $\alpha \equiv \arctan(r_i/R_i)$. The boundary conditions are very different in each of the limits $R_i \to \infty$, $i = 1, 2, 3$ or $\rho \to \infty$ (the latter requiring that both r_i, $R_i \to \infty$ simultaneously). We note that the breakup process has only been treated quantatively within a collinear model of chemical reaction; a major topic of research is the development of methods for treating breakup in full dimensionality. Henceforth, we restrict our discussion to collisions well below the breakup energy threshold. Computational methods that have been applied include algebraic variational methods and direct numerical solution of the discretized partial differential form of the Schrödinger equation. For scattering, the most widely used variational principles are based on the generalized Newton and the Kohn functionals, and they are *stationary* rather than extremal. They are typically employed using basis set expansions, with the stationary condition leading to linear systems of inhomogeneous algebraic equations. Either the basis functions explicitly include terms to satisfy the various asymptotic boundary conditions or the variational functional itself explicitly includes them through the appearance of "causal" or "anticausal" Green functions.

A major advance in dealing with the problem of coordinates in reactive scattering was the introduction of a practical method for eliminating the need to solve simultaneously for the dynamics in all three arrangements. This is done by introducing ad hoc, localized negative imaginary potentials (NIPs) designed to absorb the wave function in regions that are not of direct interest. This permits one to solve the Schrödinger equation only in the region leading from reactants to the one transition state of interest. Both a time-dependent and time-independent version of the approach exist. It is also possible to resolve state-to-state reactive scattering amplitudes at any energy contained sufficiently in the $t = 0$ wave packet. The approach yields results in quantitative agreement with those obtained by the more traditional methods. We note that in the traditional numerical integration approach, one is usually faced with solving the equations subject to so-called "two-point boundary conditions." One condition is that the solution be regular at the origin, and the others are that the solution have incident waves only in the initial arrangement and that there be only outgoing scattered waves in all possible product arrangements. Since the amplitudes for scattering into the various possible final states are unknown (and are what one is trying to compute), one only knows the *form* of the scattering boundary conditions and not their numerical values. Consequently, it is necessary to generate enough linearly independent solutions of the differential equations to form linear combinations that possess the correct asymptotic behavior. The general procedure is to expand the total wave function in internal basis states, giving rise to coupled ordinary differential equations for the expansion coefficients. If there are N basis functions in the expansion, there result N coupled second order differential equations, having in general $2N$ linearly independent solutions. At most, only N of these can be regular at the origin, leading to the necessity of solving the coupled differential equations N separate times for these N sets of linearly independent regular solutions. The expansion in the N basis functions yields an $N \times N$ Hamiltonian matrix operator, and propagating each N component regular solution vector involves doing N^2 multiplications at each step of the propagation. Since this is done N times, there is a total of N^3 multiplications at each step, for an overall scaling of MN^3, where M is the number of steps needed to escape the scattering interaction. This must be redone at each separate energy E for which one desires scattering information. By comparison, solving the time-dependent Schrödinger equation formally, one gets

$$\chi(t) = e^{-iHt/\hbar}\chi(0). \tag{95}$$

A popular way to evaluate Eq. (95) is to take t short enough that one can approximate $\exp(-iHt/\hbar)$ as the product $\exp(-iVt/2\hbar)\exp(-iH_0t/\hbar)\exp(-iVt/2\hbar)$ (called the "symmetric split operator" approximation). Then the exponential operators are evaluated in the representation in which they are diagonal (the coordinate representation for the potential and the momentum representation and internal eigenstates for the reference Hamiltonian H_0). The major computational effort is that of transforming between these two representations. For the scattering distance and its canonical momentum, this is typically done by Fast Fourier Transform, which scales as $M \log_2 M$, where M is the number of grid points in the discrete Fourier transform. The other basis dependence is at most N^2 (and can be as low as $N^{3/2}$ in the rotating frame approach). The computational effort then scales more slowly with N than the time-independent method discussed above, but more rapidly with M. Also, one must do this at each time step until the scattering is completed. To avoid having to follow the continued time evolution of the portion of $\chi(t)$ that has already escaped the range of the interaction causing the scattering, one may analyze it for the scattering information and then absorb it beyond the analysis region with an NIP. Finally, the time-dependent method automatically yields scattering information at all energies suffiently contained in $\chi(0)$, and $\chi(t)$ automatically satisfies the proper scattering boundary conditions.

A particularly powerful and promising wave packet approach makes use of the fact that while having a term $-iA$, where A is a positive semi-definite function, added to the Hamiltonian absorbs the wave packet with a time history given by $-iA\chi(t)$, if one records this time evolution information and changes the sign, $+iA\chi(t)$ then constitutes a *source* that feeds the wave back into the region where A is nonzero. This makes it possible to decouple the dynamics in all regions that can be defined using appropriate dividing surfaces from transition state theory. We write the set of *uncoupled* regional time-dependent Schrödinger equations as

$$i\hbar\frac{\partial\chi_r}{\partial t} = (H - iA_r)\chi_r, \tag{96}$$

$$i\hbar\frac{\partial\chi_2}{\partial t} = \left(H - i\sum_{j=1}^{K} A_j\right)\chi_2 + iA_r\chi_r, \tag{97}$$

$$i\hbar\frac{\partial\chi_p}{\partial t} = H\chi_p + iA_p\chi_2,\ p = 1, \ldots, K. \tag{98}$$

Here, $\chi_r(t = 0)$ is the initial packet; H is the full Hamiltonian, and $A_r,\ A_p,\ p = 1, \ldots, K$ are the functions that determine in what regions of configuration space sinks and sources are located. The absorber $-iA_r$ is located just beyond the dividing surface associated with the initial (reactant) arrangement (just inside the so-called "strong interaction" region, 2). The $-iA_p$ are located just beyond the dividing surface associated with arrangement $p,\ p = 1, \ldots, K$ (including the initial arrangement). Then it is easily seen that the dynamics of χ_r occurs *only* in the reactant arrangement, up to just within the reactant dividing surface. The dynamics of $\chi_2(t)$ occurs solely within the strong interaction region, until it is completely absorbed by the various $-iA_p$ located just outside the respective p arrangement dividing surface. The dynamics of the χ_p wave packet occurs solely in the region outside the pth dividing surface. We note that if we add all the equations, the result is

$$i\hbar\left[\sum_{p=1}^{K}\chi_p + \chi_r + \chi_2\right] = H\left[\sum_{p=1}^{K}\chi_p + \chi_r + \chi_2\right], \tag{99}$$

so that the *exact* solution of the time-dependent Schrödinger equation is

$$\chi(t) \equiv \sum_{p=1}^{K}\chi_p + \chi_r + \chi_2. \tag{100}$$

Notice that while we must solve for $\chi_r(t)$ and $\chi_2(t)$, they are nonzero for a shorter length of time than the total overall collision, and they are nonzero only in limited regions of configuration space. Therefore, computing them is much less work than solving for $\chi(t)$ directly and requires less storage. In addition, once we have recorded the time evolution of the $-iA_p\chi_2$, $p = 1, \ldots, K$, we only need to solve for the final arrangements of immediate interest, and these are done completely separately (they are totally uncoupled between two regions $p \neq p'$). We can also do the calculations whenever computing resources are most available and least costly. We also can use whatever coordinates are optimum in each separate region p. We point out that a time-independent version of Eqs. (96)–(98) has been developed which appears very robust. Essentially, it results from a half Fourier time-to-energy transform of Eq. (95), to yield [in analogy with Eq. (14)]

$$(E - H)\xi = \frac{i}{2\pi}\chi(0), \qquad (101)$$

or in causal solution form,

$$\xi^+(E) = \frac{i}{2\pi(E - H + i\epsilon)}\chi(0). \qquad (102)$$

Note that $\xi^+(E)$ is *not* the usual definite energy, causal solution of the Schrödinger equation (known as the Lippmann-Schwinger scattering solution). However, in certain well-defined regions of configuration space [namely, on the "target side" of the initial wave packet, $\chi(0)$], it is proportional to the Lippmann-Schwinger solution, $\Psi^+(E)$, satisfying

$$\Psi_k^+(E) = \phi_k(E) + \frac{1}{E - H + i\epsilon}V\phi_k(E). \qquad (103)$$

Here, $\phi_k(E)$ is the solution of the unperturbed Schrödinger equation,

$$H_0\phi_k(E) = E\phi_k(E), \qquad (104)$$

where, as usual,

$$H = H_0 + V, \qquad (105)$$

with V being the interaction responsible for causing the scattering. If we express $\chi(0)$ in terms of the complete set $\{\phi_k(E)\}$,

$$\chi(0) = \int_{-\infty}^{\infty} dkC(k)\phi_k(E), \qquad (106)$$

$$C(k) = \frac{1}{2\pi}\langle\phi_k(E)\,|\,\chi(0)\rangle, \qquad (107)$$

$$\langle\phi_k(E)\,|\,\phi_{k'}(E)\rangle = 2\pi\delta(k - k'). \qquad (108)$$

One then finds that on the target side of $\chi(0)$, in configuration space,

$$\xi^+(E) = \frac{mC(k)}{\hbar^2 k}\Psi_k^+(E). \qquad (109)$$

In general, if there are internal degrees of freedom and different possible chemical arrangements, one obtains

$$\xi^+_{rn_r}(E) = \frac{mC_r\left(k_{n_r}\right)}{\hbar^2 k_{n_r}}\Psi^+_{rk_{n_r}}(E), \qquad (110)$$

where $C_r(k_{n_r})$ characterizes the initial packet located in arrangement r, with initial internal quantum numbers n_r and relative kinetic energy $\frac{\hbar^2 k^2_{n_r}}{2m}$, and $\Psi^+_{rk_{n_r}}(E)$ is the complete Lippmann-Schwinger solution, including all possible final arrangements. The state $\xi^+_{rn_r}$ can be similarly split into localized pieces satisfying uncoupled time-independent dynamical equations analogous to Eqs. (96)–(98).

A particularly convenient observation is that, just like the homogeneous Schrödinger equation, a variational expression for the S matrix at energy E can be derived, based on Eqs. (101), (102), and (110). The result for a general scattering amplitude, connecting state r, n_r at total energy E with final state p, n_p at energy E is the robust expression

$$S\left(pn_pk_{n_p}\,\middle|\,rn_rk_{n_r}\right) = \frac{i\hbar^2\sqrt{k_{n_p}k_{n_r}}}{m(2\pi)^2C_r\left(k_{n_r}\right)C^*_p\left(k_{n_p}\right)}\left\langle\chi(pn_p\,|\,t = 0)\,\middle|\,\frac{1}{E - H + i\epsilon}\,\middle|\,\chi(rn_r\,|\,t = 0)\right\rangle. \qquad (111)$$

The modulus squared of this S matrix element is the probability of the indicated state-to-state process, and $\chi(pn_p\,|\,t = 0)$ is a wave packet located in the product region of configuration space, outside the collision region. Clearly, this is equal to a known factor times $\langle\chi(pn_p|t = 0)|\xi^+_{rn_r}(E)\rangle$. It should be clear that knowledge of the full Green function, $1/(E - H + i\epsilon) \equiv G^+(E)$, provides all one needs to know to calculate the state resolved S matrix element at the energy E.

It is worthwhile to examine how one can implement Eq. (111) computationally. A particularly effective way to do this is to expand $1/(E - H + i\epsilon)$ in an appropriate polynomial basis. A popular one is the Chebychev polynomials, denoted T_n. However, they require an argument bounded by ±1, while the exact spectrum of H is unbounded. However, in practice, one introduces a finite dimensional matrix representation of H, and this can be normalized so that it possesses no eigenvalues larger in magnitude than 1. This leads to an expansion of the form

$$\frac{1}{E - H + i\epsilon} = \sum_n g_n^+(E_{norm})T_n(H_{norm}), \qquad (112)$$

where $-1 \leq E_{norm} \leq +1$; a similar bound holds for the eigenvalues of H_{norm}; and $g_n^+(E_{norm})$ is a simple, analytically known function that explicitly builds in the causal behavior of $G^+(E)$. Computation of the S matrix then requires evaluating

$$\begin{aligned}\frac{1}{E - H + i\epsilon}\chi(rn_r \mid t = 0)\\ = \sum_n g_n^+(E_{norm})T_n(H_{norm})\chi(rn_r \mid t = 0). \qquad (113)\end{aligned}$$

Note that all of the dependence on the energy E is contained in the analytical coefficients, g_n^+. Defining "Krylov vectors"

$$\eta_n \equiv T_n(H_{norm})\chi(rn_r \mid t = 0), \qquad (114)$$

we note that they are *totally independent of E*, and

$$\frac{1}{E - H + i\epsilon}\chi(rn_r \mid t = 0) = \sum_n g^+(E_{norm})\eta_n. \qquad (115)$$

Therefore, once the $\{\eta_n\}$ are calculated that have nonzero overlap with $\chi(pn_p \mid t = 0)$, one has *all* the information needed to compute the S matrix elements at any other energy E contained in the initial packet. The T_n satisfy a simple recursion relation, so that

$$\eta_0 = \chi(rn_r \mid t = 0), \qquad (116)$$

$$\eta_1 = H_{norm}\eta_0, \qquad (117)$$

$$\eta_n = 2H_{norm}\eta_{n-1} - \eta_{n-2}, \; n \geq 2. \qquad (118)$$

If one includes any absorbing potentials, then the recursion is modified by an additional damping factor. Clearly,

$$\begin{aligned}S\big(pn_p k_{n_p} \,\big|\, rn_r k_{n_r}\big) = \frac{i\hbar^2\sqrt{k_{n_r}k_{n_p}}}{m(2\pi)^2 C_r\big(k_{n_r}\big)C_p^*\big(k_{n_p}\big)}\\ \times \sum_n g_n^+(E_{norm})\langle\chi(pn_p|t = 0)|\eta_n\rangle. \qquad (119)\end{aligned}$$

We also remark that one can easily derive a similar Chebychev expansion of $\exp(-iHt/\hbar)\chi(0)$. Then the g_n^+ are replaced by analytical functions of time, *but exactly the same η_n appear as in the energy dependent Green function expansion.* The bottom line is that results are obtained with virtually no additional effort for any energy sufficiently contained in the initial packet. Note also that if one uses the same expansion for $G^+(E)$ in Eq. (103) for the scattered wave portion of the Lippmann-Schwinger solution, one obtains $\tilde{\eta}_n(E) \equiv T_n(H_{norm})V\phi_k(E)$, which must be recalculated at every new energy!

C. Rigorous Quantum Transition State Theory

The above discussion shows, as one might have expected, that knowledge of $G^+(E)$ is tantamount to having solved, subject to causal boundary conditions, the Schrödinger equation for any possible initial and final states and arrangements. This suggests that $G^+(E)$ cannot only yield state-to-state probability amplitudes, but also the sum over all possible initial states in arrangement r and all possible final states in product arrangement p, denoted as $N_{pr}(E)$. This quantity is known as the "cummulative reaction probability," and it is central to an arbitrarily accurate, rigorous quantum transition state theory. This should also not be surprizing in light of the fact that by use of absorbing potentials located at various dividing surfaces, one can isolate the dynamics occurring in the strong interaction region. One then expects that the total strong interaction Green function for Eq. (97),

$$G_2(E) = \frac{1}{E - H + i\sum_{j=1}^{K} A_j} \qquad (120)$$

should be able to provide $N_{pr}(E)$ without having to carry out *any* dynamics outside of region 2. This has indeed been proved to be true, and it is found that

$$\begin{aligned}N_{pr}(E) &= \text{trace}\big[4A_r^{1/2}G_2^*(E)A_pG_2A_r^{1/2}\big]\\ &= \text{trace}P_{pr}(E), \qquad (121)\end{aligned}$$

where arrangement r can be any possible initial arrangement, p can be any possible final arrangement, and one computes the trace of the matrix product. The matrix $P_{pr}(E)$ is Hermitian, and its eigenvalues are interpreted as probabilities of reaction to all possible final p states from all possible initial r states. The matrix elements can be computed using *any* convenient, sufficiently complete basis and need not have any relation to asymptotic states in any of the arrangements. This provides a rigorous quantum transition state theory.

SEE ALSO THE FOLLOWING ARTICLES

Atomic and Molecular Collisions • Collision-Induced Spectroscopy • Electron Spin Resonance • Infrared Spectroscopy • Kinetics (Chemistry) • Ligand Field Concept • Microwave Molecular Spectroscopy • Quantum Mechanics • Quantum Theory

BIBLIOGRAPHY

Bader, R. (1970). “An Introduction to the Electronic Structure of Atoms and Molecules,” Clarke, Irwin, and C., Toronto.

Daudel, R., Pullman, A., Salem, L., and Veillard, A., eds. (1980, 1981, 1982). “Quantum Theory of Chemical Reactions,” Vols. I–III, Reidel, Dordrecht.

Dirac, P. A. M. (1958). “Quantum Mechanics,” Oxford Univ. Press, London.

Schatz, G. C., and Ratner, M. A. (1993). “Quantum Mechanics in Chemistry,” Prentiss Hall, Englewood Cliffs, NJ.

Simons, J., and Nichols, J. (1997). “Quantum Mechanics in Chemistry,” Oxford Univ. Press, New York.

Wyatt, R. E., and Zhang, J. Z. H., eds. (1996). “Dynamics of Molecules and Chemical Reactions,” Dekker, New York.

Quantum Chromodynamics (QCD)

Taizo Muta
Hiroshima University

GLOSSARY

Baryon Type of hadron. The baryon family includes the proton, neutron, and other particles whose eventual decay products include the proton. Baryons are composed of three-quark combinations.

Boson A particle that obeys Bose–Einstein statistics and has zero or integral spin. Unlike fermions, bosons are not conserved in number. They can be generated or destroyed singly, rather than in particle–antiparticle pairs.

Fermion A particle that obeys Fermi–Dirac statistics; a half-integer-spin particle.

Feynman diagrams Schematic representations of mathematical expressions for predicting the interaction of particles, in which lines represent the path of a particle and vertices represent particle interactions.

Gluon A massless particle that carries the strong force from one quark to another. Gluons can also interact among themselves and form particles consisting of only gluons bound together (glueballs).

Hadron Any particle of the largest family of elementary particles; they interact with each other through strong interactions, usually produce additional hadrons in a collision at high energy, and are roughly spherical.

Lepton Member of the family of weakly interacting particles, which includes the electron, muon, tau, and their associated neutrinos and antiparticles. Leptons are acted upon not by the strong force, but by the electroweak and gravitational forces.

Singularity A point in space-time where the space-time curvature becomes infinite.

THE STRONG interaction was discovered in early 1930s as nuclear forces which bind the nucleons together to form nucleus. Since then it has been a major research subject to explore the theory of strong interactions. In 1947 the first meson, the pion, was discovered experimentally as the mediator of nuclear forces. Through the 1960s the number of "elementary particles" including nucleons and mesons increased rapidly. Those elementary particles associated to strong interactions were called hadrons. In

1964 Gell-Mann and Zweig proposed that hadrons are composed of quarks. It was then natural to look for the dynamics obeyed by quark systems as the origin of strong interactions.

I. WHY IS QCD NEEDED? A HISTORICAL SURVEY

In order to obtain experimental information on quark dynamics it seems to be the most sensible to probe the inside of hadrons by applying a beam of structureless particles (i.e., leptons). For the study of the hadronic structure we need much higher energies and larger momentum transfers to obtain higher resolution. The first series of such experiments to probe the structure of the proton was initiated in the 1960s at SLAC (Stanford Linear Accelerator Center) and the process was called deep inelastic electron–proton scattering. In 1969 Bjorken reported the scaling property of the structure functions in deep inelastic electron–nucleon scatterings. This scaling is called *Bjorken scaling*, for which it is claimed that structure functions in the deep inelastic region depend only on the ratio q^2/ν rather than on two independent variables q^2 and ν, where q^2 and ν are the 4-momentum transfer squared and energy transfer of electrons, respectively.

Bjorken scaling implies that the constituents of hadrons look almost free and pointlike deep inside the nucleon when observed with high spatial resolution. These free, pointlike constituents were named *partons*. If one accepts the parton idea, the dynamics governing the parton system should have the property that the interaction between partons becomes weaker at shorter distances. The partons were later identified with quarks since experimentally it was suggested that their quantum numbers such as charges and spins were practically the same as those of quarks.

Right after the proposal of the parton model all the known quantum field theories were surveyed as possible candidates for quark dynamics. Almost all of them were shown not to enjoy the above-mentioned property that the interaction between quarks gets weaker at short distances. The exception was the non-Abelian gauge field theory, which was originally introduced by Yang and Mills. ’t Hooft, Gross, Wilczek, and Politzer found that the non-Abelian gauge field theory satisfied the desired property, which is now called *asymptotic freedom*. Soon after it was shown that non-Abelian gauge field theory was the only theory which exhibited the asymptotic freedom among the known theories in four-dimensional spacetime. Therefore the dynamics governing quark systems is to be found among non-Abelian gauge field theories. The non-Abelian gauge field theories are generated by symmetries described by a noncommutative algebra. This means that quark systems are required to have an extra symmetry associated with the non-Abelian gauge field. In the meantime three facts suggested that quarks must have a new quantum number called color and exhibit the color symmetry. We discuss these observations below.

A. The Problem of Constructing the Baryon Wave Functions

As an example, we consider the pion–nucleon resonance Δ^{++}, which is of spin 3/2 and is made of three u-quarks. If we consider the $J_3 = 3/2$ state with J_3 the third component of the total angular momentum for the Δ^{++} system, we find that all three u-quarks must have spins aligned up since relative orbital angular momentum are required to vanish for the lowest state in three-quark systems. Thus Δ^{++} state with $J_3 = 3/2$ is given by

$$\left|\Delta^{++}, J_3 = \frac{3}{2}\right\rangle = |u\uparrow, u\uparrow, u\uparrow\rangle, \tag{1}$$

where the arrow represents the spin aligned up. But this assignment is not acceptable because it contradicts the Pauli exclusion principle, according to which fermions cannot occupy the same state.

A possible way out of this difficulty may be to consider higher orbital angular momenta for the quarks. This, however, spoils the success in the prediction of baryon magnetic moments based on the S-wave three quarks. Hence we prefer to keep quarks in S-states. Then we are forced to assume the exixtence of hidden degrees of freedom for quarks, *color*, in order to distinguish three quarks which are otherwise identical. We need at least three different colors to discriminate these three quarks. It is then easy to construct the totally antisymmetric state for Δ^{++} in place of Eq. (1),

$$\left|\Delta^{++}, J_3 = \frac{3}{2}\right\rangle = \varepsilon_{ijk}|u^i\uparrow, u^j\uparrow, u^k\uparrow\rangle\ , \tag{2}$$

where indices i, j, k represent the quark colors and ε_{ijk} is the totally antisymmetric tensor (the repeated indices are summed). The same argument applies to other brayon states and the difficulty is now circummvented with the introduction of the extra color degree of freedom for quarks.

Since we do not observe the color degrees of freedom directly, we may assume that hadronic phonomena are unaltered under the exchange of colors. The symmetry group corresponding to the color degrees of freedom may be chosen from the Lie groups and we adopt SU(3) for the color symmetry. A single quark state is then assigned to the fundamental triplet, **3**, of SU(3). The state (2) is then a singlet, **1**, of SU(3).

B. Nonobservation of Isolated Quarks

Since we have no experimental evidence of hadrons which carry color quantum numbers in an explicit manner, in constructing the hadronic state out of quarks we have to pick out a singlet representation in the decomposition of the product of three triplets into irreducible representations, i.e.,

$$\mathbf{3} \otimes \mathbf{3} \otimes \mathbf{3} = \mathbf{1} \oplus \mathbf{8} \oplus \mathbf{8} \oplus \mathbf{10},$$

for the baryon state, and

$$\mathbf{3} \otimes \mathbf{3}^* = \mathbf{1} \oplus \mathbf{8},$$

for the meson state. In analogy with Eq. (2), the meson color singlet state is taken to be

$$|M\rangle = \frac{1}{3}\delta_{ij}|q^i\bar{q}^j\rangle. \tag{3}$$

The fact that color is not directly observable can be stated in a different way: physical phenomena are invariant under the color transformation. Hence the color SU(3) has to be an exact symmetry. According to this principle, all hadrons are required to be in the singlet of the color SU(3). Other states with explicit color degrees of freedom are color nonsinglets and should not be explicitly observed. It is worth noting that among the many low-lying configurations of quarks, only $q\bar{q}$ and qqq states can belong to the color singlet. Thus, the postulate of the nonobservability of colored states is essentially the same as that of the *quark confinement*, which requires that quarks be confined to the inside of hadrons and are not observed as isolated states. It should be noted that this postulate is just a kinematical constraint to eliminate colored states. There is, however, a hope that quark confinement may be a natural dynamical consequence of the strong interaction.

C. Discrepancy between Prediction and Experiment on Decay Rates for $\pi^0 \to 2\gamma$ and Total Cross Sections of $e^+e^- \to$ Hadrons

The total decay rate of $\pi^0 \to 2\gamma$ can be calculated through the lowest order Feynman diagrams, and is obtained as

$$\Gamma(\pi^0 \to 2\gamma) = N_\mathrm{c}^2\left(Q_\mathrm{u}^2 - Q_\mathrm{d}^2\right)\frac{\alpha^2 m_{\pi^0}^3}{64\pi^3 F_\pi^2}, \tag{4}$$

where N_c is the number of color degrees of freedom, Q_u and Q_d are the u- and d-quark charges in units of the proton charge e, m_{π^0} is the neutral pion mass, $\alpha = e^2/4\pi$, and F_π is the pion decay constant for $\pi \to \mu\nu$ decays ($F_\pi = 91$ MeV). Substituting $N_\mathrm{c} = 3$, $Q_\mathrm{u} = 2/3$ and $Q_\mathrm{d} = -1/3$ we have $\Gamma(\pi^0 \to 2\gamma) = 7.6$ eV, which is in perfect agreement with the experimental data $\Gamma(\exp) = 7.48 \pm 0.33$ eV. Note that the theoretical prediction with $N_\mathrm{c} = 1$ (no color degree of freedom), $\Gamma(\pi^0 \to 2\gamma) = 0.84$ eV, is far from explaining the data.

For $e^+ + e^-$ annihilations at very high energies the total cross section $\sigma(e^+e^- \to \text{hadrons})$ is given by

$$\sigma(e^+e^- \to \text{hadrons}) = \frac{4\pi\alpha^2}{3s}N_\mathrm{c}\sum_{i=1}^{N_\mathrm{f}} Q_i^2, \tag{5}$$

where s is the center-of-mass total energy squared of the e^+e^- system, Q_i is the charge of the ith quark, and N_f is the number of flavors of quarks which may contribute to the process. Comparision of Eq. (5) with the data strongly support $N_\mathrm{c} = 3$. If there is no color degree of freedom, the prediction of Eq. (5) is much smaller than experimental data.

These facts strongly support the idea that quarks should have an extra quantum number, color. In 1970s the idea of the extra quantum number, color, for quarks was established and the theory of quark dynamics was found to be a non-Abelian gauge theory with SU(3) symmetry that was named quantum chromodynamics (QCD).

The non-Abelian gauge field in QCD mediates color interactions between quarks. It is called the *gluon*. Gluons carry color charges and hence interact with each other even in the absence of quarks. This property of gluons is an essential ingredient for having asymptotic freedom. When one considers massless gluons, serious infrared divergences exist in QCD and may be the origin of quark confinement. The QCD has the property of the asymptotic freedom at short distances, while it has the possibility of quark confinement at long distances.

According to the property of asymptotic freedom of QCD, one may safely use perturbation theory to discuss short-distance processes. This approach in QCD is usually referred to as *perturbative* QCD. Perturbative QCD has been applied to many physical processes with the help of operator product expansion (OPE) and renormalization group equations, such as electron–nucleon scattering, e^+e^- annihilation, heavy quarkonia (J/ψ's and Υ's) decays, photon–photon scattering as a subprocess of $e^+e^- \to e^+e^-X$ (X represents unobserved hadrons), production of jets from quarks and gluons, the Drell–Yan process $NN \to l^+l^-X$, inclusive e^+e^- annihilation $e^+e^- \to \text{hadrons} + X$, large-$p_T$ hadron reactions, and multijets in e^+e^- annihilations. These applications, together with the later analysis of further experimental data, gave strong support to QCD.

II. PRINCIPLES OF QCD

QCD is a gauge field theory, which is based on the gauge principle. The gauge principle is the requirement that the theory be invariant under the local gauge transformation.

The symmetry group in QCD is the color SU(3) group, which is a noncommutative (non-Abelian) group. A gauge theory which is invariant under the non-Abelian gauge group is also called Yang–Mills theory.

We consider the fermion $\psi(x)$ with mass m (the quark field, to be more specific) which belongs to the N-dimensional fundamental representation of the group G. Thus the field $\psi(x)$ has N components: $\psi_i(x)$, $i = 1, 2, \ldots, N$. The group G in QCD is the color SU(3) group and so $N = 3$. The Lagrangian for the free fermion field $\psi(x)$ is given by

$$\mathcal{L} = \bar{\psi}_i \big(i\gamma^\mu \partial_\mu - m\big)\psi_i. \tag{6}$$

The Lagrangian is invariant under the transformation

$$\psi' = U_{ij}\psi_j, \qquad U = \exp(-iT^a\theta^a), \tag{7}$$

where θ^a ($a = 1, \ldots, 8$) are the transformation parameters and T^a are the SU(3) generators, which are subject to the commutation relations

$$[T^a, T^b] = if^{abc}T^c, \tag{8}$$

where the summation on the repeated indices should be understood as usual and f^{abc} are the structure constants of the SU(3) group. For θ^a depending on x, the Lagrangian (6) is no longer invariant under the transformation (7). The Lagranian may be made invariant under the non-Ablian local gauge transformation (7) if the derivative in Eq. (6) is replaced by the covariant derivative,

$$D_\mu = \partial_\mu - igT^aA^a_\mu, \tag{9}$$

where A^a_μ are the gauge fields and g is a constant representing the coupling strength between ψ and A^a_μ. In component form Eq. (9) reads

$$(D_\mu)_{ij} = \delta_{ij}\partial_\mu - igT^a_{ij}A^a_\mu, \tag{10}$$

where T^a_{ij} is the representation of T^a in the fundamental representation. We now replace the Lagrangian (6) by

$$\begin{aligned}\mathcal{L} &= \bar{\psi}_i\,\big(i\gamma^\mu(D_\mu)_{ij} - m\delta_{ij}\big)\,\psi_i \\ &= \bar{\psi}\big(i\gamma^\mu D_\mu - m\big)\psi.\end{aligned} \tag{11}$$

The new Lagrangian (11) is invariant under the non-Ablian local gauge transformation (7) provided $A^a_\mu(x)$ obeys the transformation rule

$$T^aA'^a_\mu = U\bigg(T^aA^a_\mu - \frac{i}{g}U^{-1}\partial_\mu U\bigg)U^{-1}. \tag{12}$$

We now show that the Lagranian (11) is invariant under the local gauge transformation. Note that

$$\begin{aligned}(D_\mu\psi)' &= \big(\partial_\mu - igT^aA'^a_\mu\big)\psi' \\ &= U\big(\partial_\mu + U^{-1}\partial_\mu U - igU^{-1}T^aUA'^a_\mu\big)\psi.\end{aligned} \tag{13}$$

Because A^a_μ satisfies the transformation rule (12), we have

$$(D_\mu\psi)' = U(D_\mu\psi). \tag{14}$$

Thus $\bar{\psi}D_\mu\psi$ is invariant under the non-Abelian local gauge transformation (12) and hence the Lagrangian (11) is also invariant.

The Lagrangian (11) describes the fermion $\psi(x)$ in interaction with the gauge fields $A^a_\mu(x)$. It is still to be supplemented by a kinetic term consisting purely of the gauge fields $A^a_\mu(x)$. Inspired by the example of electrodynamics, we may try the form

$$-\frac{1}{4}\big(\partial_\mu A^a_\nu - \partial_\nu A^a_\mu\big)(\partial^\mu A^{a\nu} - \partial^\nu A^{a\mu}). \tag{15}$$

However, such a term is not SU(3) gauge invariant. It is not difficult to prove that $F^a_{\mu\nu}F^{a\mu\nu}$ is invariant with

$$F^a_{\mu\nu} \equiv \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + gf^{abc}A^b_\mu A^c_\nu. \tag{16}$$

Finally we arrive at the general form of the Lagrangian which is invariant under the non-Abelian local gauge transformations (7) and (12),

$$\mathcal{L} = -\frac{1}{4}F^a_{\mu\nu}F^{a\mu\nu} + \bar{\psi}\big(i\gamma^\mu D_\mu - m\big)\psi. \tag{17}$$

It should be noted that in the above Lagrangian there exists only one arbitrary parameter g owing to gauge invariance. This universal constant g is called the *gauge coupling constant*. We derived the above Lagrangian based on the gauge principle. Of course it should still be Lorentz invariant and invariant under space and time reversal. Unfortunately, the Lagrangian (17) is not unique, in that we may add to it other terms of higher power of $F^a_{\mu\nu}$ and ψ within the requirement of local gauge invariance, Lorentz invariance, and invariance under space inversion and time reversal. The additional requirement of renormalizability may eliminate all the irrelevant terms and fix the Lagrangian (17) in a unique way.

After fixing the Lagrangian we are ready to move on to the quantization of the theory under consideration. The quantization can be conveniently performed in the Feynman functional-integral formalism. For simplicity we consider a system consisting only of a neutral scalar field $\phi(x)$ with mass m. The Green function is given by the functional integral

$$\langle 0|T[\hat{\phi}(x_1)\cdots\hat{\phi}(x_n)]|0\rangle = \frac{\int[d\phi]\phi(x_1)\cdots\phi(x_n)\exp(iS)}{\int[d\phi]\exp(iS)}, \tag{18}$$

where S is the classical action,

$$S = \int d^4x\mathcal{L}. \tag{19}$$

Here $\mathcal{L}$ is the classical Lagrangian density:

$$\mathcal{L} = \frac{1}{2}\left(\partial^\mu\phi\partial_\mu\phi - m^2\phi^2\right) - V(\phi), \tag{20}$$

with $V(\phi)$ the part of the Lagrangian representing the self-coupling of the field ϕ. The Green function (18) may be reexpressed in another form if we introduce an external source function $J(x)$ and an artificial source term $\phi(x)J(x)$ in the functional integral,

$$Z[J] = \int [d\phi] \exp\left\{ i \int d^4x(\mathcal{L} + \phi J)\right\}. \tag{21}$$

Here $Z(J)$ is a functional of $J(x)$. We define the functional differentiation by

$$\frac{\delta Z[J(x)]}{\delta J(y)} = \lim_{\varepsilon\to 0} \frac{Z[J(x) + \varepsilon\delta(x-y)] - Z[J(x)]}{\varepsilon}. \tag{22}$$

According to the above definition, we have

$$\frac{\delta^n Z[J]}{\delta J(x_1)\cdots\delta J(x_n)} = i^n \int [d\phi]\,\phi(x_1)\cdots\phi(x_n) \exp\left\{ i \int d^4x(\mathcal{L} + \phi J)\right\}. \tag{23}$$

Hence we obtain

$$\langle 0|T[\hat{\phi}(x_1)\cdots\hat{\phi}(x_n)]|0\rangle = \frac{(-i)^n}{Z[0]} \left.\frac{\delta^n Z[J]}{\delta J(x_1)\cdots\delta J(x_n)}\right|_{J=0}. \tag{24}$$

The functional integral $Z(J)$ thus generates all the Green functions. In this sense $Z(J)$ is called the *generating functional* for Green functions.

A straightforward application of Eq. (21) to the case of gauge fields suggests that the generating functional for gauge fields A^a_μ is given by

$$Z[J] = \int [dA] \exp\left\{ i \int d^4x\left(\mathcal{L} + A^a_\mu J^{a\mu}\right)\right\}, \tag{25}$$

where $\mathcal{L}$ is given by Eq. (11) and $[A]$ is the shorthand notation for

$$\prod_{\mu,a}[dA^a_\mu]. \tag{26}$$

In Eq. (25), $\mathcal{L}$ and $[A]$ are gauge invariant. The action $S = \int d^4x\mathcal{L}$ is also gauge invariant. In the following we set $A'^a_\mu = A^{(\theta)a}_\mu$ in order to emphasize its dependence on θ^a. Starting with a fixed A^a_μ, we obtain a set of $A^{(\theta)a}_\mu$ by applying to A^a_μ all the transformations $U(\theta)$ belonging to group G. According to the above invariance, the action S is constant for all $A^{(\theta)a}_\mu$ in this subset and the functional integral $Z[0]$ on this subset of $A^{(\theta)a}_\mu$ diverges, as the region of the integral is infinite. Hence it is sensible to integrate only once on such $A^{(\theta)a}_\mu$ that belongs to the subset and to factor out a divergent constant. After this mamipulation is done with the following restriction on A^a_μ,

$$G^\mu A^a_\mu = B^a, \tag{27}$$

we have an expression for $Z[J]$,

$$Z[J] = \int [dA] \det M_G \prod_{a,x} \delta\left(G^\mu A^a_\mu(x) - B^a(x)\right) \times \exp\left\{ i \int d^4x\left(\mathcal{L} + A^a_\mu J^{a\mu}\right)\right\}. \tag{28}$$

Since $B^a(x)$ is arbitrary, we may average $Z[J]$ over $B^a(x)$ in the sense of the functional integral, i.e., we integrate $Z[J]$ on $B^a(x)$ with a suitable weight, which we choose to be

$$\exp\left\{ -(i/2\alpha) \int d^4x(B^a(x))^2\right\}, \tag{29}$$

where α is an arbitrary constant. Hence we obtain

$$Z[J] = \int [dA] \det M_G \times \exp\left\{ i \int d^4x\left(\mathcal{L} - \frac{1}{2\alpha}\left(G^\mu A^a_\mu\right)^2 + A^a_\mu J^{a\mu}\right)\right\}. \tag{30}$$

In this way we succeeded in exponentiating the constraint. The resulting exponent is the so-called gauge-fixing term with gauge parameter α. The following are examples of explicit expressions for the matrix M_G:

1. *Coulomb gauge* $G^\mu = (0, \nabla)$:

$$(M_G(x,y))^{ab} = (\delta^{ab}\nabla^2 - gf^{abc}\mathbf{A}^c\cdot\nabla)\delta^4(x-y). \tag{31}$$

2. *Lorentz (covariant) gauge* $G^\mu = \partial^\mu$:

$$(M_G(x,y))^{ab} = \left(\delta^{ab}\Box - gf^{abc}\partial^\mu A^c_\mu\right)\delta^4(x-y). \tag{32}$$

3. *Axial gauge* $G^\mu = n^\mu$ (here n^μ is a spacelike constant 4-vector):

$$(M_G(x,y))^{ab} = (\delta^{ab}n\cdot\partial - gf^{abc}n\cdot A^c)\delta^4(x-y). \tag{33}$$

4. *Temporal gauge* $G^\mu = (1, 0, 0, 0)$:

$$(M_G(x,y))^{ab} = \left(\delta^{ab}\partial_0 - gf^{abc}A^c_0\right)\delta^4(x-y). \tag{34}$$

With the generating functional $Z[J]$ given by Eq. (30) the quantization program for gauge fields is completed.

The functional-integral quantization has been successfully performed for scalar and gauge fields in Eqs. (21) and (30). The case of fermion fields needs special care and the quantization is performed by the use of the Grassmann number, which is a set of anticommuting numbers ψ_j $(j = 1, \ldots, N)$,

$$\{\psi_i, \psi_j\} = 0. \tag{35}$$

Thus the generating functional including fermion fields is

$$Z[J, \eta, \bar{\eta}] = \int [dA][d\psi][d\bar{\psi}] \det M_G \times \exp\left\{ i \int d^4x \left(\mathcal{L} - (1/2\alpha)\left(\partial^\mu A^a_\mu\right)^2 + A^a_\mu J^{a\mu} + \bar{\psi}\eta + \bar{\eta}\psi\right)\right\}, \tag{36}$$

where η and $\bar{\eta}$ are anticommuting source functions for fermion fields $\bar{\psi}$ and ψ, with $\bar{\psi} = \psi^\dagger \gamma_0$. According to the anticommuting property of fermion fields and source functions, we have to pay special attention to the sign associated with $\eta(x)$ when we define fermion Green functions. For example, the fermion two-point Green function should be defined in the following way:

$$\langle 0|T[\hat{\psi}_\alpha(x)\bar{\hat{\psi}}_\beta(y)]|0\rangle = \frac{(-i)^2}{Z[0,0,0]} \left.\frac{\delta^2 Z[J,\eta,\bar{\eta}]}{\delta\bar{\eta}_\alpha(x)\delta(-\eta_\beta(y))}\right|_{J=\bar{\eta}=\eta=0}. \tag{37}$$

This rule of doing the functional differentiation with $-\eta(x)$ should always be kept in mind in defining fermion Green functions.

The following formula for the integral over a Grassman algebra will be useful in the following, where $A(x, y)$ is a certain complex function:

$$\int [d\psi][d\bar{\psi}] \exp\left\{ \int d^4x\, d^4y \bar{\psi}(x) A(x, y)\psi(y)\right\} = \det A. \tag{38}$$

III. GENERAL FRAMEWORK OF QCD. PERTURBATIVE REGIME

A. Perturbation Theory

In order to develop the perturbation theory it is most convenient to use the covariant gauge. In this case, however, $\det M_G$ given by Eq. (32) depends on A^a_μ and also on the gauge coupling constant g and hence a simple perturbative expansion of Eq. (36) is not allowed. For this purpose we need to exponentiate $\det M_G$ and regard it as a part of the effective Lagrangian. Up to an irrelevant factor, $\det M_G$ is represented by, according to Eq. (38),

$$\det M_G = \int [d\chi][d\chi^*] \exp\left\{ -i \int d^4x\, d^4y \chi^a(x)^* \times (M_G(x, y))^{ab} \chi^b(y)\right\}, \tag{39}$$

where M_G is given by Eq. (32) and $\chi^a(x)$ is a complex fictitious field obeying the Grassmann algebra and belonging to the adjoint representation of the gauge group $\mathrm{G} = \mathrm{SU}(3)$. The field $\chi^a(x)$ is called the *Faddeev–Popov ghost*, as it has the strange property that it is fermonic as well as bosonic. The exponent of the integrand in Eq. (39) can be rewritten by doing an integration by parts such that

$$\int d^4x\, d^4y \chi^a(x)^* (M_G(x, y))^{ab} \chi^b(y) = -\int d^4x (\partial^\mu \chi^a(x))^* D^{ab}_\mu \chi^b(x), \tag{40}$$

where D^{ab}_μ is the covariant derivative in the adjoint representation [note that $(T^a)_{bc} = -if^{abc}$],

$$D^{ab}_\mu = \delta^{ab}\partial_\mu - gf^{abc}A^c_\mu. \tag{41}$$

We insert Eq. (39) together with Eq. (40) into Eq. (36) to obtain the generating functional

$$Z[J, \xi, \xi^*, \eta, \bar{\eta}] = \int [dA][d\chi][d\chi^*][d\psi][d\bar{\psi}] \times \exp\left\{ i \int d^4x (\mathcal{L} + AJ + \chi^*\xi + \xi^*\chi + \bar{\psi}\eta + \bar{\eta}\psi)\right\}, \tag{42}$$

where ξ^a and ξ^{a*} are source functions (Grassmann numbers) for the ghosts, and AJ, $\chi^*\xi$, and $\xi^*\chi$ are shorthand notations for

$$AJ = A^a_\mu J^{a\mu}, \quad \chi^*\xi = \chi^{a*}\xi^a, \quad \xi^*\chi = \xi^{a*}\chi^a. \tag{43}$$

Here $\mathcal{L}$ is an effective quantum Lagrangian which includes the effect of $\det M_G$,

$$\mathcal{L} = \mathcal{L}_{\mathrm{G}} + \mathcal{L}_{\mathrm{GF}} + \mathcal{L}_{\mathrm{FP}} + \mathcal{L}_{\mathrm{F}}, \tag{44}$$

$$\mathcal{L}_{\mathrm{G}} = -\frac{1}{4}F^a_{\mu\nu}F^{a\mu\nu},$$

$$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + gf^{abc}A^b_\mu A^c_\nu, \tag{45}$$

$$\mathcal{L}_{\mathrm{GF}} = -\frac{1}{2\alpha}\left(\partial^\mu A^a_\mu\right)^2, \tag{46}$$

$$\mathcal{L}_{\mathrm{FP}} = (\partial^\mu \chi^{a*})D^{ab}_\mu \chi^b, \tag{47}$$

$$\mathcal{L}_{\mathrm{F}} = \bar{\psi}^i\left(i\gamma^\mu D^{ij}_\mu - m\delta^{ij}\right)\psi^j. \tag{48}$$

The indices on the Lagrangians (45)–(48) stand for "gauge," "gauge fixing," "Faddeev–Popov," and "fermion" terms, respectively. The Lagrangian (44) forms the basis of quantum chromodynamics.

In the following we will develop the perturbation theory of quantum chromodynamics and derive the Feynman rules for it. For this purpose we split up the Lagrangian (44) into a free part $\mathcal{L}_0$ and an interaction part $\mathcal{L}_1$,

$$\mathcal{L} = \mathcal{L}_0 + \mathcal{L}_1 \tag{49}$$

with

$$\mathcal{L}_0 = \mathcal{L}_0^{\mathrm{G}} + \mathcal{L}_0^{\mathrm{FP}} + \mathcal{L}_0^{\mathrm{F}}, \tag{50}$$

$$\begin{aligned}\mathcal{L}_0^{\mathrm{G}} &= -\frac{1}{4}\big(\partial_\mu A_\nu^a - \partial_\nu A_\mu^a\big)(\partial^\mu A^{a\nu} - \partial^\nu A^{a\mu})\\ &\quad - \frac{1}{2\alpha}\big(\partial^\mu A_\mu^a\big)^2,\end{aligned} \tag{51}$$

$$\mathcal{L}_0^{\mathrm{FP}} = (\partial^\mu \chi^{a*})\big(\partial_\mu \chi^a\big), \tag{52}$$

$$\mathcal{L}_0^{\mathrm{F}} = \bar{\psi}\big(i\gamma^\mu \partial_\mu - m\big)\psi. \tag{53}$$

As Eq. (52) is of the form of the Lagrangian for massless charged scalar fields, we recognize that the Faddeev–Popov ghost is bosonic although it is fermionic owing to its nature as a Grassmann number. The remaining part of the Lagrangian $\mathcal{L}$ after subtracting $\mathcal{L}_0$ is the interaction Lagrangian $\mathcal{L}_1$,

$$\begin{aligned}\mathcal{L}_1 &= \mathcal{L}_1(A^a, \chi^a, \chi^{a*}, \psi, \bar{\psi})\\ &= -\frac{g}{2} f^{abc}\big(\partial_\mu A_\nu^a - \partial_\nu A_\mu^a\big) A^{b\mu} A^{c\nu}\\ &\quad - \frac{g^2}{4} f^{abe} f^{cde} A_\mu^a A_\nu^b A^{c\mu} A^{d\nu}\\ &\quad - g f^{abc}(\partial^\mu \chi^{a*})\chi^b A_\mu^c + g\bar{\psi} T^a \gamma^\mu \psi A_\mu^a.\end{aligned} \tag{54}$$

Then Eq. (42) can be rewritten in the following form:

$$\begin{aligned}&Z[J, \xi, \xi^*, \eta, \bar{\eta}]\\ &\quad = \exp\bigg\{ i \int d^4x \mathcal{L}_1\\ &\qquad \times \bigg(\frac{\delta}{i\delta J^{a\mu}}, \frac{\delta}{i\delta \xi^{a*}}, \frac{\delta}{i\delta(-\xi^a)}, \frac{\delta}{i\delta\bar{\eta}}, \frac{\delta}{i\delta(-\eta)} \bigg) \bigg\}\\ &\qquad \times Z_0[J, \xi, \xi^*, \eta, \bar{\eta}],\end{aligned} \tag{55}$$

where Z_0 is a generating functional for free fields,

$$Z_0[J, \ldots] = Z_0^{\mathrm{G}}[J] Z_0^{\mathrm{FP}}[\xi, \xi^*] Z_0^{\mathrm{F}}[\eta, \bar{\eta}], \tag{56}$$

$$Z_0^{\mathrm{G}}[J] = \int [dA] \exp\bigg\{ i \int d^4x \big(\mathcal{L}_0^{\mathrm{G}} + AJ\big) \bigg\}, \tag{57}$$

$$\begin{aligned}Z_0^{\mathrm{FP}}[\xi, \xi^*] &= \int [d\chi][d\chi^*]\\ &\quad \times \exp\bigg\{ i \int d^4x \big(\mathcal{L}_0^{\mathrm{FP}} + \chi^*\xi + \xi^*\chi\big) \bigg\},\end{aligned} \tag{58}$$

$$\begin{aligned}Z_0^{\mathrm{F}}[\eta, \bar{\eta}] &= \int [d\psi][d\bar{\psi}]\\ &\quad \times \exp\bigg\{ i \int d^4x \big(\mathcal{L}_0^{\mathrm{F}} + \bar{\psi}\eta + \bar{\eta}\psi\big) \bigg\}.\end{aligned} \tag{59}$$

In order to obtain $Z[J, \ldots]$ perturbatively we first calculate $Z_0[J, \ldots]$ for the gluon, Faddeev–Popov ghost, and quark, respectively. To do this we reexpress the free Lagrangian by doing an integration by parts,

$$\begin{aligned}\mathcal{L}_0^{\mathrm{G}} &= -\frac{1}{2} A_\mu^a K^{ab\mu\nu} A_\nu^b,\\ K_{\mu\nu}^{ab} &= \delta^{ab}\bigg(-g_{\mu\nu}\Box + \bigg(1 - \frac{1}{\alpha}\bigg) \partial_\mu \partial_\nu \bigg),\end{aligned} \tag{60}$$

$$\mathcal{L}_0^{\mathrm{FP}} = \chi^{a*} K^{ab} \chi^b, \qquad K^{ab} = \delta^{ab}\Box, \tag{61}$$

$$\mathcal{L}^{\mathrm{F}} = -\bar{\psi}\Lambda\psi, \qquad \Lambda = -i\gamma^\mu \partial_\mu + m. \tag{62}$$

If we denote the inverses of $K_{\mu\nu}^{ab}$, K^{ab}, and Λ by $D_{\mu\nu}^{ab}$, D^{ab}, and S, respectively, we have

$$\int d^4z K_{\mu\lambda}^{ac}(x - z) g^{\lambda\rho} D_{\rho\nu}^{cb}(z - y) = \delta^{ab} g_{\mu\nu} \delta^4(x - y), \tag{63}$$

$$\int d^4z K^{ac}(x - z) D^{cb}(z - y) = \delta^{ab}\delta^4(x - y), \tag{64}$$

$$\int d^4z \Lambda(x - z) S(z - y) = \delta^4(x - y). \tag{65}$$

These functions $D_{\mu\nu}^{ab}$, D^{ab}, and S are propagators of the gluon, Feddeev–Popov ghost and quark, respectively. Solving the conditions (63)–(65) for Fourier coefficients, we find

$$D_{\mu\nu}^{ab}(x) = \delta^{ab} \int \frac{d^4k}{(2\pi)^4} \frac{e^{-ik\cdot x}}{k^2 + i\varepsilon} \bigg(g_{\mu\nu} - (1 - \alpha)\frac{k_\mu k_\nu}{k^2} \bigg), \tag{66}$$

$$D^{ab}(x) = \delta^{ab} \int \frac{d^4k}{(2\pi)^4} \frac{-1}{k^2 + i\varepsilon} e^{-ik\cdot x}, \tag{67}$$

$$S(x) = \int \frac{d^4p}{(2\pi)^4} \frac{1}{m - \not{p}} e^{-ip\cdot x}. \tag{68}$$

We can now perform the functional integrations of A_μ^a, χ, χ^*, ψ, and $\bar{\psi}$ in Eqs. (57)–(59), respectively. We obtain, apart from irrelevant constant multiples,

$$Z_0^{\mathrm{G}}[J] = \exp\bigg\{ \frac{i}{2} \int d^4x\, d^4y\, J^{a\mu}(x) D_{\mu\nu}^{ab}(x - y) J^{b\nu}(y) \bigg\}, \tag{69}$$

$$Z_0^{\mathrm{FP}}[\xi, \xi^*] = \exp\bigg\{ i \int d^4x\, d^4y \xi^a(x)^* D^{ab}\ (x - y)\xi^b(y) \bigg\}, \tag{70}$$

$$Z_0^{\mathrm{F}}[\eta, \bar{\eta}] = \exp\bigg\{ i \int d^4x\, d^4y \bar{\eta}(x) S(x - y) \eta(y) \bigg\}. \tag{71}$$

We insert Eqs. (69)–(71) into Eq. (56) and use Eq. (55) to generate the perturbation series,

$$Z[J,\ldots] = \left\{1 + i\int d^4x\,\mathcal{L}_1\left(\frac{\delta}{i\delta J^{a\mu}(x)},\ldots\right) + \cdots\right\} Z_0[J,\ldots]. \tag{72}$$

We can calculate Green functions order by order with the use of Eq. (72). For example, we can calculate two-point Green functions (propagator), three-point Green functions (vertex), etc. Accordingly we have the Feynman rules for the Lagrangian of QCD shown in Tables I and II.

Now that we have settled the Lagrangian for QCD and established the Feynman rules for it, we are free to make perturbative calculations of cross sections for an arbitrary quark–gluon process. In general, however, the loop contributions to quark–gluon processes generate divergences. The divergences are properly taken care of by the renormalization program which we will describe in the following.

Renormalization is a technique for subtracting the divergences appearing in the loop calculations. Before substracting these divergences the divergent integrals should

TABLE I QCD Feynman Rules

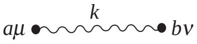
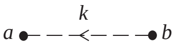
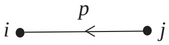
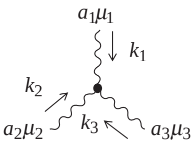
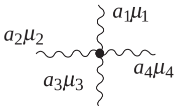
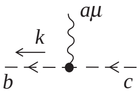
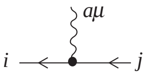
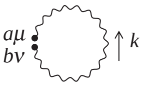
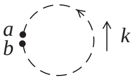
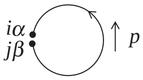
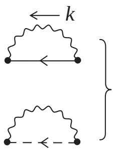
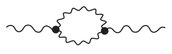
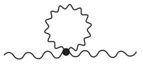
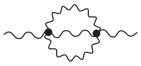
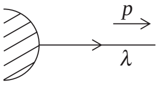
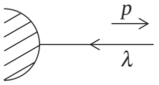
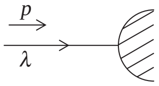
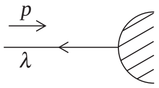
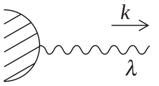
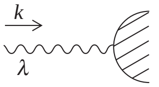
Propagators		
Gluons A		$\delta_{ab}\dfrac{d_{\mu\nu}(k)}{k^2}$
Ghosts χ		$\delta_{ab}\dfrac{-1}{k^2}$
Quarks ψ		$\delta_{ij}\dfrac{1}{m-\not{p}}$
Vertices		
Three-Gluon		$-igf^{a_1a_2a_3}V_{\mu_1\mu_2\mu_3}(k_1,k_2,k_3)$
Four-gluon		$-g^2W^{a_1\ldots a_4}_{\mu_1\ldots\mu_4}$
Gluon–ghost		$-igf^{abc}k_\mu$
Gluon–quark		$g\gamma_\mu T^a_{ij}$
Loops		
Gluon		$\displaystyle\int\frac{d^4k}{(2\pi)^4i}\delta^{ab}\delta^{\mu\nu}$
Ghost		$\displaystyle-\int\frac{d^4k}{(2\pi)^4i}\delta^{ab}$
Quark		$\displaystyle-\int\frac{d^4p}{(2\pi)^4i}\delta^{ij}\delta^{\alpha\beta}$
Gluon–quark		$\displaystyle\int\frac{d^4k}{(2\pi)^4i}$
Gluon–ghost		
Symmetry factors		
$\frac{1}{2!}$  ,	$\frac{1}{2!}$  ,	$\frac{1}{3!}$ 

TABLE II Additional Rules for Transition Matrix Elements

Mass-shell condition for external lines		$p^2 = m^2$
Fermions (quarks)		
Outgoing fermion		$\bar{u}_\lambda(p)$
Outgoing antifermion		$v_\lambda(p)$
Incoming fermion		$u_\lambda(p)$
Incoming antifermion		$\bar{v}_\lambda(p)$
Vector fields (gluons)		
Outgoing vector		$\varepsilon_\lambda^\mu(k)$
Incoming vector		$\varepsilon_\lambda^\mu(k)$

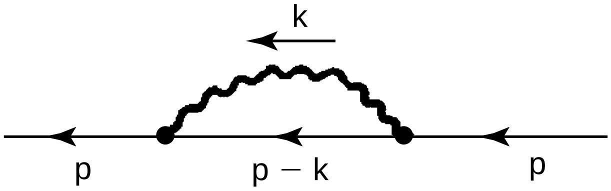


FIGURE 1

be made tentatively finite by introducing a suitable convergence device. This procedure is generically called *regularization*. Regularization is a purely mathematical procedure which has no physical consequences; accordingly, it is not a unique procedure. We have a variety of the regularization schems, such as (1) the cutoff method, (2) the Pauli–Villars regulator method, (3) analytic regularization, (4) lattice regularization, and (5) dimensional regularization. Since in dimensional regularization the regularized theory is kept Lorentz invariant, gauge invariant, and unitary, in this sense dimensional regularization is the most suitable for gauge theories. We will present this method in this article.

In order to explain the dimensional regularization scheme, we take a specific example of a divergent integral, the quark self-energy part $\Sigma_{ij}(p)$. Its relation to the quark propagator $\tilde{S}_{ij}(p)$ is given by

$$\tilde{S}_{ij}(p) = \delta_{ij}/(m - \not{p} - \Sigma(p)). \qquad (73)$$

Following the Feynman rules for QCD, we can obtain an expression for the quark self-energy part to order g^2 (see Fig. 1),

$$\Sigma(p) = g^2 C_F \int \frac{d^4k}{(2\pi)^4 i} \frac{\gamma_\mu(m + \not{p} - \not{k})\gamma^\mu}{k^2(m^2 - (p-k)^2)}. \qquad (74)$$

The above expression is obtained in Feynman gauge $\alpha = 1$, and $C_F = (N^2 - 1)/2N$, $N = 3$. The four-dimensional integral in Eq. (74) is linearly divergent, as can be easily seen by simple power counting in k, i.e.,

$$\int d^4k \frac{\not{k}}{k^2 k^2} \sim \lim_{K\to\infty} K. \qquad (75)$$

The divergence comes from the high-momentum region $|k| \to \infty$. This divergent integral may be made convergent by reducing the number of multiple integrals. For example, Eq. (74) would be finite if the space-time were two dimensional. This fact is the basic idea of dimensional regularization. Where we keep the space-time dimension D lower than four and replace the divergent four-dimensional integral by a convergent D-dimensional one. By making momentum integrations explicitly, we obtain an analytic expression as a function of the dimension D. We make the analytic continuation in D in this expression. Then the original divergence will show up as a pole at $D = 4$ in the above analytic expression.

We summarize the convensions for dimensional regularization here:

1. The D-dimensional space-time has the metric $g^{\mu\nu} = (+, -, \ldots, -)$.
2. $\mathrm{Tr}[I] = 4$ in the space of the gamma matrices.
3. The integral measure is $\int d^Dk/(2\pi)^D$.
4. γ_5 is an object which satisfies $\{\gamma_5, \gamma^\mu\} = 0$.

It is worth noting here that the gauge coupling constant g is no longer dimensionless for arbitrary space-time dimensions, $\dim[g] = 2 - D/2$. Thus we introduce a mass scale μ by hand and rewrite the gauge coupling constant g in the following way:

$$g = g_0 \mu^{2-D/2}, \qquad (76)$$

where g_0 is the dimensionless gauge coupling constant. We can finally obtain the expression for the quark self-energy part in D-dimensional space-time for the covariant gauge with α arbitrary,

$$\begin{aligned}\Sigma(p) &= \alpha \frac{2C_F g^2}{(4\pi)^{D/2}} \not{p}\,(-p^2)^{D/2-2}(D-1)B \\ &\quad \times \left(\frac{D}{2}, \frac{D}{2}\right)\Gamma\left(2 - \frac{D}{2}\right) \\ &= \alpha \frac{g_0^2}{(4\pi)^2} C_F \not{p}\left(\frac{1}{\varepsilon} - \gamma + 1 - \ln\frac{-p^2}{4\pi\mu^2}\right) + O(\varepsilon),\end{aligned} \qquad (77)$$

where we defined a new parameter ε by $\varepsilon = (4 - D)/2$. We can find that there is a ploe in Eq. (77) which is relevant to the condition when the arbitrary dimension D is four.

In the renormalization program divergences in Green functions are subtracted by redefining the fields, the coupling constant g, and the mass parameter m in the original Lagrangian. It is important to note here that the way of eliminting divergences in perturbation theory is not unique because there exists an ambiguity in defining the divergent piece of the Green function. This ambiguity eventually leads to ambiguity in the finite piece of the Green function. In order to remove this ambiguity, we have to specify how we define the divergent piece which will be subtracted out in the renormalization process. The prescription for subtracting divergences in Green functions is called the *renormalizaion scheme*. The scheme in which we eliminate only the pole term $1/\varepsilon$ in the dimensionally regularized expression of the Green functions is called *minimal subtraction* (MS). We see in Eq. (77) that the pole term is usually accompanied by the natural constant γ and $\ln 4\pi$ in the combination $1/\varepsilon - \gamma + \ln 4\pi$, thus the scheme to eliminate the whole of $\underline{1/\varepsilon - \gamma + \ln 4\pi}$ is called *modified minimal subtraction* ($\overline{\mathrm{MS}}$).

The basic Lagrangian for QCD is given in Eq. (44). It is more convenient to rewrite the FP ghost term in the form

$$\mathcal{L}_{\mathrm{FP}} = i\left(\partial^\mu \chi_1^a\right) D_\mu^{ab} \chi_2^b. \tag{78}$$

Hence the QCD Lagrangian is taken as

$$\begin{aligned}\mathcal{L} = &-\frac{1}{4}F^a_{\mu\nu}F^{a\mu\nu} - \frac{1}{2\alpha}\left(\partial^\mu A^a_\mu\right)^2 \\ &+ i\left(\partial^\mu\chi_1^a\right)D_\mu^{ab}\chi_2^b + \bar\psi^i\left(i\gamma^\mu D_\mu^{ij} - m\delta^{ij}\right)\psi^j. \end{aligned}\tag{79}$$

The above Lagrangian may be decomposed into a free and an interaction part, $\mathcal{L}_0$ and $\mathcal{L}_1$, such that

$$\mathcal{L} = \mathcal{L}_0 + \mathcal{L}_1, \tag{80}$$

$$\begin{aligned}\mathcal{L}_0 = &-\frac{1}{4}\left(\partial_\mu A^a_\nu - \partial_\nu A^a_\mu\right)\left(\partial^\mu A^{a\nu} - \partial^\nu A^{a\mu}\right) - \frac{1}{2\alpha}\left(\partial^\mu A^a_\mu\right)^2 \\ &+ i\left(\partial^\mu\chi_1^a\right)\left(\partial_\mu\chi_2^a\right) + \bar\psi^i\left(i\gamma^\mu\partial_\mu - m\right)\psi^i, \end{aligned}\tag{81}$$

$$\begin{aligned}\mathcal{L}_1 = &-\frac{g}{2}f^{abc}\left(\partial_\mu A^a_\nu - \partial_\nu A^a_\mu\right)A^{b\mu}A^{c\nu} \\ &-\frac{g^2}{4}f^{abe}f^{cde}A^a_\mu A^b_\nu A^{c\mu}A^{d\nu} \\ &-igf^{abc}\left(\partial^\mu\chi_1^a\right)\chi_2^b A^c_\mu + g\bar\psi^i T^a_{ij}\gamma^\mu\psi^j A^a_\mu. \end{aligned}\tag{82}$$

To eliminate the divergences in loop corrections to the Green functions, we redfine the fields A^a_μ, χ_1^a, χ_2^a, and ψ by

$$A^a_\mu = Z_3^{1/2}A^a_{r\mu}, \qquad \chi^a_{1,2} = \tilde Z_3^{1/2}\chi^a_{1,2r}, \qquad \psi = Z_2^{1/2}\psi_r, \tag{83}$$

and the parameters g, α, and m by

$$g = Z_g g_r, \qquad \alpha = Z_3\alpha_r, \qquad m = Z_m m_r, \tag{84}$$

where the constants Z_3, $\tilde Z_3$, Z_2, Z_m, and Z_g are called renormalization constants. Inserting Eqs. (83) and (84) into Eq. (79), we obtain

$$\mathcal{L} = \mathcal{L}_{r0} + \mathcal{L}_{r1} + \mathcal{L}_{\mathrm{C}}, \tag{85}$$

where $\mathcal{L}_{r0}$ and $\mathcal{L}_{r1}$ are precisely equal to $\mathcal{L}_0$ and $\mathcal{L}_1$ if the quantities A^a_μ, $\chi^a_{1,2}$, ψ, g, α, and m are replaced by the renormalized ones, $A^a_{r\mu}$, $\chi^a_{1,2r}$, ψ_r, g_r, α_r, and m_r. Then $\mathcal{L}_{\mathrm{C}}$ is given by

$$\begin{aligned}\mathcal{L}_{\mathrm{C}} = &-(Z_3 - 1)\frac{1}{4}\left(\partial_\mu A^a_{r\nu} - \partial_\nu A^a_{r\mu}\right)\left(\partial^\mu A^{a\nu}_r - \partial^\nu A^{a\mu}_r\right) \\ &+(\tilde Z_3 - 1)i\left(\partial^\mu\chi^a_{1r}\right)\left(\partial_\mu\chi^a_{2r}\right) \\ &+(Z_2 - 1)\bar\psi^i_r\left(i\gamma^\mu\partial_\mu - m_r\right)\psi^i_r \\ &-Z_2(Z_m - 1)m_r\bar\psi^i_r\psi^i_r \\ &-\left(Z_gZ_3^{3/2} - 1\right)\frac{1}{2}g_r f^{abc}\left(\partial_\mu A^a_{r\nu} - \partial_\nu A^a_{r\mu}\right)A^{b\mu}_r A^{c\nu}_r \\ &-\left(Z_g^2Z_3^2 - 1\right)\frac{1}{4}g_r f^{abe}f^{cde}A^a_{r\mu}A^b_{r\nu}A^{c\mu}_r A^{d\nu}_r \\ &-\left(Z_g\tilde Z_3 Z_3^{1/2} - 1\right)ig_r f^{abc}\left(\partial^\mu\chi^a_{1r}\right)\chi^b_{2r}A^c_{r\mu} \\ &+\left(Z_gZ_2Z_3^{1/2} - 1\right)g_r\bar\psi^i_r T^a_{ij}\gamma^\mu\psi^j_r A^a_{r\mu}. \end{aligned}\tag{86}$$

If we define four new renormalization constants by

$$\begin{aligned} Z_1 \equiv Z_gZ_3^{3/2}, \qquad Z_4 \equiv Z_g^2Z_3^2\tilde Z_1 \equiv Z_g\tilde Z_3Z_3^{1/2}, \\ Z_{1F} \equiv Z_gZ_2Z_3^{1/2}. \end{aligned}\tag{87}$$

We can rewrite Eq. (86) in the form

$$\begin{aligned}\mathcal{L}_{\mathrm{C}} = &(Z_3 - 1)\frac{1}{2}A^{a\mu}_r\delta_{ab}\left(g_{\mu\nu}\partial^2 - \partial_\mu\partial_\nu\right)A^{b\nu}_r \\ &+(\tilde Z_3 - 1)\chi^a_{1r}\delta_{ab}(-i\partial^2)\chi^b_{2r} \\ &+(Z_2 - 1)\bar\psi^i_r\left(i\gamma^\mu\partial_\mu - m_r\right)\psi^i_r \\ &-Z_2(Z_m - 1)m_r\bar\psi^i_r\psi^i_r \\ &-(Z_1 - 1)\frac{1}{2}g_r f^{abc}\left(\partial_\mu A^a_{r\nu} - \partial_\nu A^a_{r\mu}\right)A^{b\mu}_r A^{c\nu}_r \\ &-(Z_4 - 1)\frac{1}{4}g_r f^{abe}f^{cde}A^a_{r\mu}A^b_{r\nu}A^{c\mu}_r A^{d\nu}_r \\ &-(\tilde Z_1 - 1)ig_r f^{abc}\left(\partial^\mu\chi^a_{1r}\right)\chi^b_{2r}A^c_{r\mu} \\ &+(Z_{1F} - 1)g_r\bar\psi^i_r T^a_{ij}\gamma^\mu\psi^j_r A^a_{r\mu}. \end{aligned}\tag{88}$$

The term $\mathcal{L}_{\mathrm{C}}$ is called the counter-term Lagrangian, and is used to subtract the divergences. The renormalization constants Z_3, $\tilde{Z}_3$, Z_2, Z_m, and Z_g should be determined by adjusting the counter terms so as to cancel overall divergences appearing in higher order Feynman amplitudes.

B. Renormalization Group Method

According to the renormalization program, we substract all the divergences from the Green functions systematically order by order in perturbation theory. In the subtraction procedure there exists an arbitrariness of how to define a divergent piece in a Green function, i.e., how much of the finite piece is to be subtracted together with the infinity. This arbitrariness results in a variety of renormalization schemes. On the other hand, in subtracting the divergences we inevitably introduce an arbitrary mass scale μ, which is called the *renormalization scale.* The renormalization scale μ is entirely arbitrary and remains in the finite part of the Green functions, thus leaving an arbitrariness for the renormalized Green functions after the subtraction of divergences.

Due to this arbitrariness we have many possible expressions for one physical quantity depending on the choice of the renormalization scheme and scale. These different expressions are connected by a finite renormalization. Since they describe a unique physical phenomenon and they have to be equivalent. In other words, physical quantities such as S-matrix elements are invariant under a finite renormalization.

The renormalized coupling constants g_r and m_r depend on the renormalization scale μ at which the subtraction procedure is defined. Writing this dependence explicitly, we have

$$\begin{aligned} g_r(\mu) &= Z_g(\mu)^{-1} g, \\ m_r(\mu) &= Z_m(\mu)^{-1/2} m. \end{aligned} \tag{89}$$

The renormalized coupling constants $g_r(\mu)$ and $g_r(\mu')$ which are obtained through two different subtraction procedures characterized by the renormalization scales μ and μ', respectively, are related to each other by

$$g_r(\mu') = z_g(\mu', \mu) g_r(\mu), \tag{90}$$

where the finite renormalization $z_g(\mu', \mu)$ is given by

$$z_g(\mu', \mu) = Z_g(\mu)/Z_g(\mu'). \tag{91}$$

In the same way we have

$$m_r(\mu') = z_m(\mu', \mu) m_r(\mu), \tag{92}$$

where $z_g(\mu', \mu)$ is defined by

$$z_m(\mu', \mu) = [Z_m(\mu)/Z_m(\mu')]^{1/2}. \tag{93}$$

Equation (90) defines a set of finite renormalization $\{z_g(\mu', \mu)\}$ for varying renormalization scales μ' and μ. The finite renormalization (90) can be regarded as a transformation. This set of transformations possesses group properties. This group is Abelian. A similar property is possessed by $\{z_m(\mu', \mu)\}$, which is also an Abelian group.

We have yet another kind of finite renormalization which applies to Green functions. For simplicity, let us consider here the truncated connected Green function for n gluon legs $\tilde{G}_n^{\mathrm{tc}}(p, g, m)$, which is defined by

$$\begin{aligned} &(2\pi)^4 \delta^4(p_1 + \cdots + p_n) \tilde{G}_2(p_1) \cdots \tilde{G}_2(p_n) \tilde{G}_n^{\mathrm{tc}}(p, g, m) \\ &\quad = \int d^4x_1 \cdots d^4x_n e^{-i(p_1\cdot x_1 + \cdots + p_n\cdot x_n)} G_n^{\mathrm{c}}(x_1, \ldots, x_n), \end{aligned} \tag{94}$$

where $G_n^{\mathrm{c}}(x_1, \ldots, x_n)$ is the connected Green function given by

$$G_n^{\mathrm{c}}(x_1, \ldots, x_n) = (-i)^{n-1} \left.\frac{\delta^n W[J, g, m]}{\delta J(x_1) \ldots \delta J(x_n)}\right|_{J=0}, \tag{95}$$

and $W[J, g, m]$ is the generating functional for the connected Green functions, which is related to the generating functional $Z[J, g, m]$ for the Green functions G_n through the relation

$$Z[J, g, m] = e^{iW[J,g,m]}. \tag{96}$$

Since

$$W[J, g, m] = W_r[J_r, g_r, m_r], \tag{97}$$

and $J = Z_3^{-1/2} J_r$ (note here the redefinition of J corresponding to the field redefinition $A_\mu^a = Z_3^{1/2} A_{r\mu}^a$), we find that the renormalized connected Green function $G_{rn}^{\mathrm{c}}(x_1, \ldots, x_n)$ is related to $G_n^{\mathrm{c}}(x_1, \ldots, x_n)$ by

$$G_{rn}^{\mathrm{c}} = Z_3^{-n/2} G_n^{\mathrm{c}}. \tag{98}$$

Defining the renormalized truncated connected Green function $\tilde{G}_{rn}^{\mathrm{tc}}(p, g_r, m_r, \mu)$ through the equation

$$\begin{aligned} &(2\pi)^4 \delta^4(p_1 + \cdots + p_n) \tilde{G}_{r2}(p_1) \cdots \tilde{G}_{r2}(p_n) \\ &\quad \times \tilde{G}_{rn}^{\mathrm{tc}}(p, g_r, m_r, \mu) \\ &\quad = \int d^4x_1 \cdots d^4x_n e^{-i(p_1\cdot x_1 + \cdots + p_n\cdot x_n)} G_{rn}^{\mathrm{c}}(x_1, \ldots, x_n), \end{aligned} \tag{99}$$

we obtain

$$\tilde{G}_{rn}^{\mathrm{tc}}(p, g_r, m_r, \mu) = Z_3(\mu)^{n/2} \tilde{G}_n^{\mathrm{tc}}(p, g, m). \tag{100}$$

We introduce the renormalized Feynman amplitude $F_n(p, g_r(\mu), m_r(\mu), \mu)$ with renormalization scale μ such that

$$F_n(p, g_r(\mu), m_r(\mu), \mu)$$
$$= -i\tilde{G}^{\mathrm{tc}}_m(p, g_r(\mu), m_r(\mu), \mu). \qquad (101)$$

The finite renormalization for F_n is then given by

$$F_n(p, g_r(\mu'), m_r(\mu'), \mu')$$
$$= z_n(\mu', \mu)F_n(p, g_r(\mu), m_r(\mu), \mu), \qquad (102)$$

where the renormalization factor $z_n(\mu', \mu)$ is defined by

$$z_n(\mu', \mu) = [Z_3(\mu')/Z_3(\mu)]^{n/2}. \qquad (103)$$

Thus the change of the renormalization scale $\mu \to \mu'$ generates a finite multiplicative renormalization of the Feynman amplitudes which gives rise to an Abelian group $\{z_n(\mu', \mu)\}$.

We have obtained three sets of finite renormalizations $\{z_g(\mu', \mu)\}$, $\{z_m(\mu', \mu)\}$, and $\{z_n(\mu', \mu)\}$ which constitute Abelian groups generated by the scale change $\mu \to \mu'$. The group thus obtained is called the *renormalization group*.

The transformations (90), (92), and (102) may be regarded as functional equations for $g_r(\mu)$, $m_r(\mu)$, and $F_n(p, g_r(\mu), m_r(\mu), \mu)$ characteristic of the renormalization group. If we restrict ourselves to an infinitesimal change of the renormalization scale μ, these functional equations reduce to differential equations which correspond to the Lie differential equations in the Lie group. These differential equations are called *renormalization group equations*.

Now we derive the renormalization group equation in the MS (or $\overline{\mathrm{MS}}$) scheme. We first derive the differential equations corresponding to the functional equations (90) and (92). We keep μ fixed in Eqs. (90) and (92) and differentiate both sides of these functional equations with respect to μ'. It is, however, more convenient in the later practical calculations to start with Eq. (89) to obtain the same differential equations. We employ dimensional regularization. Then g and g_r acquire mass dimension. We isolate these mass dimensions explicitly,

$$\begin{aligned} g &= g_0\mu_0^{\varepsilon} \\ g_r &= g_R\mu^{\varepsilon}, \end{aligned} \qquad (104)$$

where $\varepsilon = (4 - D)/2$ and g_0 and g_R are dimensionless coupling constants. Here the mass scale μ_0 is fixed scale, while the mass scale μ for the renormalized coupling constant g_r is a variable parameter. The mass scale μ is in fact identified with the renormalization scale in the MS (or $\overline{\mathrm{MS}}$) scheme. Using Eq. (104), we rewrite Eq. (89) in the following form:

$$g_{\mathrm{R}}(\mu) = \left(\frac{\mu_0}{\mu}\right)^{\varepsilon} Z_g(\mu)^{-1}g_0. \qquad (105)$$

The bare parameter g and m are regarded as fixed constants, hence we have

$$\frac{dg}{d\mu} = 0, \qquad \frac{dm}{d\mu} = 0. \qquad (106)$$

According to Eq. (105), we obtain, from (106), the differential equations for the renormalized parameters g_{R} and m_{R},

$$\mu\frac{dg_{\mathrm{R}}}{d\mu} = \beta \qquad (107)$$

$$\mu\frac{dm_{\mathrm{R}}}{d\mu} = -m_{\mathrm{R}}\gamma_m, \qquad (108)$$

where we have rewritten m_r as m_{R}, i.e., $m_{\mathrm{R}} = m_r$, just to balance the notation, and β and γ_m are given by

$$\beta = -\varepsilon g_{\mathrm{R}} - \frac{\mu}{Z_g}\frac{dZ_g}{d\mu}g_{\mathrm{R}}, \qquad (109)$$

$$\gamma_m = \frac{\mu}{Z_m^{1/2}}\frac{dZ_m^{1/2}}{d\mu}. \qquad (110)$$

The quantities β and γ_m defined here are finite functions of μ since the divergences in Z_g and Z_m as $\varepsilon \to 0$ cancel out in expressions (109) and (110).

We now turn our attention to the differential equation corresponding to the functional equation (102). We note that the unrenormalized n-point Green function $\overline{G}^{\mathrm{tc}}_n(p, g, m)$ defined in Eq. (94) is independent of the renormalization scale μ as far as the bare parameters g and m are fixed, i.e.,

$$\frac{d}{d\mu}\tilde{G}^{\mathrm{tc}}_n(p, g, m)|_{g,m} = 0. \qquad (111)$$

In terms of the renormalized n-point Green function defined by Eq. (100), we reexpress Eq. (111) in the following way:

$$\frac{d}{d\mu}\big[Z_3(\mu, g, m)^{-n/2}F_n(p, g_{\mathrm{R}}(\mu, g, m),$$
$$\times\, m_{\mathrm{R}}(\mu, g, m), \mu)\big]\big|_{g,m} = 0. \qquad (112)$$

Applying the chain rule in differentiation to Eq. (112), we obtain

$$\left.\frac{\partial Z_3^{-n/2}}{\partial\mu}\right|_{g,m} F_n + Z_3^{-n/2}$$
$$\times\left(\frac{\partial}{\partial\mu} + \frac{\partial g_{\mathrm{R}}}{\partial\mu}\frac{\partial}{\partial g_{\mathrm{R}}} + \frac{\partial m_{\mathrm{R}}}{\partial\mu}\frac{\partial}{\partial m_{\mathrm{R}}}\right)F_n\bigg|_{g,m} = 0. \qquad (113)$$

Rewriting Eq. (113), we have

$$\left(\mu\frac{\partial}{\partial\mu} + \beta\frac{\partial}{\partial g_{\mathrm{R}}} - \gamma_m m_{\mathrm{R}}\frac{\partial}{\partial m_{\mathrm{R}}} - n\gamma\right)F_n = 0, \qquad (114)$$

where β, γ_m, and γ are defined respectively by

$$\beta = \mu \frac{\partial g_{\mathrm{R}}}{\partial \mu}\bigg|_{g,m}, \tag{115}$$

$$\gamma_m = -\frac{\mu}{m_{\mathrm{R}}}\frac{\partial m_{\mathrm{R}}}{\partial \mu}\bigg|_{g,m}, \tag{116}$$

$$\gamma = -\frac{\mu}{2Z_3}\frac{\partial Z_3}{\partial \mu}\bigg|_{g,m}. \tag{117}$$

It should be mention that in the MS scheme β, γ_m, and γ depend only on g_{R}:

$$\beta = \beta(g_{\mathrm{R}}), \tag{118}$$

$$\gamma_m = \gamma_m(g_{\mathrm{R}}), \tag{119}$$

$$\gamma = \gamma(g_{\mathrm{R}}). \tag{120}$$

Equation (114) together with Eqs. (118)–(120) constitute the renormalization group equation for the Green function F_n in the MS scheme. Equation (114) is called the ’t Hooft–Weinberg equation. The functions $\beta(g_{\mathrm{R}})$, $\gamma_m(g_{\mathrm{R}})$, and $\gamma(g_{\mathrm{R}})$ are called the renormalization group functions. In particular $\beta(g_{\mathrm{R}})$ goes by the name of the β-function or the Gell-Mann–Low function, and $\gamma(g_{\mathrm{R}})$ is called the anomalous dimension of the gluon field A_μ^a.

Now we can easily generalize Eq. (114) to the truncated connected Green function F_{n_G,n_F} with n_G gluon and n_F quark legs, and take the renormalized gauge constant α_{R} into account. The generalized renormalization group equation reads

$$\begin{aligned}\bigg[\mu\frac{\partial}{\partial\mu} &+ \beta(g_{\mathrm{R}},\alpha_{\mathrm{R}})\frac{\partial}{\partial g_{\mathrm{R}}} - \gamma_m(g_{\mathrm{R}},\alpha_{\mathrm{R}})m_{\mathrm{R}}\frac{\partial}{\partial m_{\mathrm{R}}}\\ &+ \delta(g_{\mathrm{R}},\alpha_{\mathrm{R}})\frac{\partial}{\partial\alpha_{\mathrm{R}}} - n_G\gamma_G(g_{\mathrm{R}},\alpha_{\mathrm{R}})\\ &- n_F\gamma_F(g_{\mathrm{R}},\alpha_{\mathrm{R}})\bigg]F_{n_G,n_F} = 0,\end{aligned} \tag{121}$$

where $g_{\mathrm{R}} = g_r\mu^{-\varepsilon}$, $g_0 = g\mu_0^{-\varepsilon}$, $\varepsilon = (4-D)/2$, $m_{\mathrm{R}} = m_r$, and $\alpha_{\mathrm{R}} = \alpha_r$. The renormalization group functions β, γ_m, δ, γ_G, and γ_F are defined by

$$\beta(g_{\mathrm{R}},\alpha_{\mathrm{R}}) = \mu\frac{\partial g_{\mathrm{R}}}{\partial\mu}\bigg|_{g,m,\alpha}, \tag{122}$$

$$\gamma_m(g_{\mathrm{R}},\alpha_{\mathrm{R}}) = -\frac{\mu}{m_{\mathrm{R}}}\frac{\partial m_{\mathrm{R}}}{\partial\mu}\bigg|_{g,m,\alpha}, \tag{123}$$

$$\delta(g_{\mathrm{R}},\alpha_{\mathrm{R}}) = \mu\frac{\partial\alpha_{\mathrm{R}}}{\partial\mu}\bigg|_{g,m,\alpha} = -2\alpha_{\mathrm{R}}\gamma_G(g_{\mathrm{R}},\alpha_{\mathrm{R}}), \tag{124}$$

$$\gamma_G(g_{\mathrm{R}},\alpha_{\mathrm{R}}) = -\frac{\mu}{2Z_3}\frac{\partial Z_3}{\partial\mu}\bigg|_{g,m,\alpha}, \tag{125}$$

$$\gamma_F(g_{\mathrm{R}},\alpha_{\mathrm{R}}) = -\frac{\mu}{2Z_2}\frac{\partial Z_2}{\partial\mu}\bigg|_{g,m,\alpha}. \tag{126}$$

Here γ_G and γ_F are called the anomalous dimensions of the gluon and quark fields, respectively.

The renormalization group functions can be calculated order by order in quantum chromodynamics. The calculation of the β-function has been performed up to four loops in the MS scheme. Here we present the expression for the β-function up to three loops:

$$\beta(g) = -\beta_0 g^3 - \beta_1 g^5 - \beta_2 g^7 + O(g^9), \tag{127}$$

where the β_i are given by

$$\beta_0 = \frac{1}{(4\pi)^2}\left(11 - \frac{2}{3}N_f\right), \tag{128}$$

$$\beta_1 = \frac{1}{(4\pi)^4}\left(102 - \frac{38}{3}N_f\right), \tag{129}$$

$$\beta_2 = \frac{1}{(4\pi)^6}\left(\frac{2857}{2} - \frac{5033}{18}N_f + \frac{325}{54}N_f^2\right). \tag{130}$$

Now we introduce the running coupling constant $\bar{g}$, which is the renormalized running coupling constant defined at the arbitrary renormalization scale μ, i.e.,

$$\mu\frac{d\bar{g}}{d\mu} = \beta(\bar{g}), \tag{131}$$

If we reexpress the scale μ by a new scale t through the relation $\mu = e^t$, the above equation becomes

$$\frac{d\bar{g}}{dt} = \beta(\bar{g}), \tag{132}$$

where the running coupling constant $\bar{g}$ can be regarded as function of t, so that it should be expressed as $\bar{g}(t)$. The integrated form of Eq. (132) is given by

$$t = \int_g^{\bar{g}(t)} \frac{dg'}{\beta(g')}. \tag{133}$$

We choose the momentum scale to be

$$e^t = \sqrt{-q^2}/\mu, \tag{134}$$

where q is the typical momentum under consideration, which is taken to be spacelike, and μ_0 is the fixed momentum scale. We insert Eq. (127) into Eq. (133) to obtain

$$t = -\frac{1}{2}\int_{g^2}^{\bar{g}^2}\frac{d\lambda}{\lambda^2}\frac{1}{\beta_0 + \beta_1\lambda + \beta_2\lambda^2 + O(\lambda^3)}, \tag{135}$$

where $g = \bar{g}(0)$. If we choose g sufficiently small, λ is also kept small since $\bar{g} < g$. Hence to this approximation we may safely truncate the perturbative series for the β-function. Keeping only the one-loop order, we have from Eq. (135)

$$t = \frac{1}{2\beta_0}\left(\frac{1}{\bar{g}^2} + \frac{1}{g^2}\right). \tag{136}$$

Hence the running coupling constant $\bar{g}$ is given by

$$\bar{g}^2 = \frac{g^2}{1 + 2\beta_0 g^2 t} = \frac{1}{\beta_0 \ln\left(-q^2/\Lambda^2\right)}, \tag{137}$$

where the new momentum scale Λ is defined by

$$\Lambda = \mu \exp\left[-1/\left(2\beta_0 g^2\right)\right]. \tag{138}$$

The momentum scale Λ is often referred to as the QCD *scale parameter* and is the only adjustable parameter in QCD except for the quark masses. In fact, the free parameter g present in the original Lagrangian is replaced by the scale parameter Λ through Eq. (138). The scale parameter Λ should be determined by comparing the QCD predictions with experimental data.

From Eq. (137), we can see that asymptotic freedom occurs if $\beta_0 > 0$. Equation (128) shows that the condition for the asymptotic freedom reads

$$N_{\mathrm{f}} < 33/2.$$

Hence quantum chromodynamics enjoys the property of asymptotic freedom in so far as the number of quark flavors is less than 16. It should be noted that, for $N_{\mathrm{f}} = 0$, i.e., for a world made up only of gluons, the coefficient β_0 is positive definite. It is the presence of quarks that can spoil asymptotic freedom. The fundamental origin of asymptotic freedom may be traced back to the existence of the three-gluon coupling term in the Lagrangian. As this term is peculiar to the Yang–Mills theory, we realize that asymptotic freedom is inherent in the nature of a Yang–Mills theory.

The formula (137) may be improved by taking into account the two-loop term, i.e., the term with the coefficient β_1 in Eq. (135). Performing the integration, we obtain

$$t = \frac{1}{2\beta_0}\left[\frac{1}{\bar{g}^2} - \frac{1}{g^2} + \frac{\beta_1}{\beta_0} \ln \frac{\bar{g}^2\left(\beta_0 + \beta_1 g^2\right)}{g^2\left(\beta_0 + \beta_1 \bar{g}^2\right)}\right]. \tag{139}$$

We rewrite (139) in the following form:

$$\frac{1}{\bar{g}^2} + \frac{\beta_1}{\beta_0} \ln \frac{\beta_0 \bar{g}^2}{1 + \beta_1 \bar{g}^2/\beta_0} = \beta_0 \ln\left(\frac{-q^2}{\Lambda^2}\right), \tag{140}$$

with the scale parameter Λ defined by

$$\Lambda = \mu e^{-1/(2\beta_0 g^2)} \left(\frac{1 + \beta_1 g^2/\beta_0}{\beta_0 g^2}\right)^{\beta_1/(2\beta_0^2)}. \tag{141}$$

Note that Eq. (141) reduces to Eq. (138) if we set $\beta_1 = 0$. Equation (140) may be solved for $\bar{g}^2$ iteratively provided that $-q^2 \gg \Lambda^2$,

$$\bar{g}^2 = \frac{1}{\beta_0 \ln\left(-q^2/\Lambda^2\right)}\left[1 - \frac{\beta_1}{\beta_0^2} \frac{\ln\ln(-q^2/\Lambda^2)}{\ln(-q^2/\Lambda^2)} + \cdots\right]. \tag{142}$$

The second term in the parentheses in the above equation represents the next to leading order, which corresponds to the two-loop effect.

We find that the renormalized coupling constant tends to be small as the relevant momentum scale grows. According to this property of asymptotic freedom, we realize that our perturbative calculation is justified for the large-momentum scale. Thus, in QCD, perturbation theory is perfectly legitimate in the large-momentum region.

C. Operator-Product Expansion

The product of fields at the same space-time point is called the composite field or composite operator. Strictly speaking, the composite operator field is not well defined if one takes a product of fields in a naive way. To make the argument simpler we shall confine ourselves to the case of the neutral scalar field $\phi(x)$.

Even for free fields we can see show that the composite operator is not well defined. Let us consider the time-ordered product of two fields $T[\phi(x)\phi(y)]$. Its vacuum expectation value is the propagator of the free field $\phi(x)$ (times $-i$),

$$\begin{aligned}\langle 0|T[\hat{\phi}(x)\hat{\phi}(y)]|0\rangle &= -i\,\Delta(x-y)\\ &= -i\int \frac{d^4p}{(2\pi)^4}\frac{e^{-ip\cdot(x-y)}}{m^2 - p^2 - i\varepsilon}.\end{aligned} \tag{143}$$

As we let $y \to x$ in Eq. (143), we see that the momentum integral on the right-hand side diverges. Furthermore, not only for the vacuum expectation value (143), but also for general Green functions $\langle 0|T[\hat{\phi}(x)\hat{\phi}(y)\hat{\phi}(x_1)\cdots\hat{\phi}(x_n)]|0\rangle$ one may show that the divergence occurs as $y \to x$. Thus the composite operator $\lim_{y\to x} T[\hat{\phi}(x)\hat{\phi}(y)]$ for free fields is not well defined. To see the situation more clearly, we perform the momentum integration in Eq. (143) explicitly:

$$\begin{aligned}\Delta(x-y) &= \frac{1}{4\pi}\delta((x-y)^2)\\ &\quad + i\frac{m}{4\pi^2}\frac{K_1\left(m\sqrt{-(x-y)^2 + i\varepsilon}\right)}{\sqrt{-(x-y)^2 + i\varepsilon}},\end{aligned} \tag{144}$$

where $K_1(z)$ is the modified Bessel function of the second kind. The right-hand side of Eq. (144) is obviously divergent for $x = y$. In the case of free fields we may find a meaningful definition of the composition field $\phi(x)^2$ by subtracting the singularity of its vacuum expectation value from the naive product of the field operators,

$$:\hat{\phi}(x)^2: = \lim_{y\to x}\{\hat{\phi}(x)\hat{\phi}(y) - \langle 0|\hat{\phi}(x)\hat{\phi}(y)|0\rangle\}. \tag{145}$$

The composite operator $:\hat{\phi}(x)^2:$ defined in this way is nothing but the normal product of free fields. For the interacting fields, this simple manipulation cannot be generalized in a straightforward manner.

For interacting fields we also have a simple argument showing that the composite operator $\phi(x)^2$ is ill defined. We take the vacuum expectation value of the composite operator $\hat{\phi}(x)^2$ and find that

$$\langle 0|\hat{\phi}(x)^2|0\rangle = \int \frac{d^3p}{(2\pi)^3 2p_0} \sum_n |\langle 0|\hat{\phi}(0)^2|p, n\rangle|^2. \quad (146)$$

Here we have inserted between the two $\hat{\phi}(x)$'s the complete set of eigenstates $|p, n\rangle$ of the four-momentum operator $\hat{P}_\mu$,

$$\hat{P}_\mu |p, n\rangle = p_\mu |p, n\rangle, \quad (147)$$

where n is a quantum number other than momentum p, labeling the eigenstates, and we have also used the translation invariance of the theory,

$$\hat{\phi}(x) = e^{i\hat{P}\cdot x}\hat{\phi}(0)e^{-i\hat{P}\cdot x}. \quad (148)$$

The complete set of states $|p, n\rangle$ as a subset and hence

$$\sum_n |\langle 0|\hat{\phi}(0)|p, n\rangle|^2 \geq |\langle 0|\hat{\phi}(0)|p, 1\rangle|^2. \quad (149)$$

Here the matrix element in Eq. (149) depends only on p^2 by covariance. In particular, the right-hand side of Eq. (149) is independent of p_μ since $p^2 = m^2$ for the one-particle state, where m is the mass of field $\phi(x)$. Therefore we have by combining Eq. (146) with Eq. (149),

$$\langle 0|\hat{\phi}(x)^2|0\rangle \geq N \int \frac{d^3p}{(2\pi)^4 2p_0}, \qquad p_0 = \sqrt{\mathbf{p}^2 + m^2}, \quad (150)$$

the right-hand side of which is divergent, and $N = |\langle 0|\hat{\phi}(0)|p, 1\rangle|^2$. Thus the composite operator $\hat{\phi}(x)^2$ in general gives rise to divergent matrix elements and is not a mathematically well-defined object.

The operator-product expansion proposed by Wilson may serve to give a meaningful definition of the composite operator. By the operator-product expansion we mean that the product of operators, say $\hat{A}(x)$ and $\hat{B}(x)$, is expanded in a series of well-defined local operators $\hat{O}_i(x)$ with singular c-number coefficients $C_i(x)$ $(i = 0, 1, 2, 3, \ldots)$,

$$\hat{A}(x)\hat{B}(y) = \sum_{i=0}^{\infty} C_i(x - y)\hat{O}_i\left(\frac{x + y}{2}\right), \quad (151)$$

where $\hat{A}(x)$ and $\hat{B}(x)$ may be the field $\hat{\phi}(x)$ or any other local operators. The local operator $\hat{O}_i(x)$ is regular in the sense that the singularity of the product $\hat{A}(x)\hat{B}(x)$ for $y = x$ is fully contained in the coefficient functions $C_i(x - y)$. In Eq. (151) we arranged each term in the order of decreasing singularity. Hence $C_0(x - y)$ is the most singular as $y \to x$, the next most singular one is $C_i(x - y)$, and so on. The operator $\hat{O}_0(x)$ is usually an identity operator. Thus the operator-product expansion serves as a means of defining the composite operator.

Now we shall derive the operator-product expansion of Eq. (151). We exemplify it in free field theories. One of the simplest examples of the operator-product expansion in free field theories is the Wick theorem applied to the time-ordered product of two free neutral scalar fields:

$$T[\phi(x)\phi(y)] = :\phi(x)\phi(y): + \langle 0|T[\phi(x)\phi(y)]|0\rangle, \quad (152)$$

where $:\phi(x)\phi(y):$ represents the normal product. Also, for free fermions we have

$$T[\psi(x)\bar{\psi}(y)] = :\psi(x)\bar{\psi}(y): + \langle 0|T[\psi(x)\bar{\psi}(y)]|0\rangle. \quad (153)$$

As remarked before, the normal product of free fields may be used to define a composite operator, as it is regular even in the limit $x \to y$. Defining the electromagnetic current $j_\mu(x)$ by the normal product for the quark fields,

$$j_\mu(x) = :\bar{\psi}(x)\gamma_\mu\psi(x):, \quad (154)$$

we derive the expansion of the product of two currents by applying the Wick theorem,

$$\begin{aligned} T[j_\mu(x)j_\nu(0)] &= -\mathrm{Tr}[\langle 0|T[\psi(0)\bar{\psi}(x)]|0\rangle\gamma_\mu\langle 0|T[\psi(x)\bar{\psi}(0)]|0\rangle\gamma_\nu] \\ &\quad + :\bar{\psi}(x)\gamma_\mu\langle 0|T[\psi(x)\bar{\psi}(0)]|0\rangle\gamma_\nu\psi(0): \\ &\quad + :\bar{\psi}(0)\gamma_\nu\langle 0|T[\psi(0)\bar{\psi}(x)]|0\rangle\gamma_\mu\psi(x): \\ &\quad + :\bar{\psi}(x)\gamma_\mu\psi(x)\bar{\psi}(0)\gamma_\nu\psi(0):. \end{aligned} \quad (155)$$

Note here that for the free quark field $\psi(x)$

$$i\langle 0|T[\psi(x)\bar{\psi}(0)]|0\rangle = S(x) = \int \frac{d^4p}{(2\pi)^4}\frac{1}{m - \not{p} - i\varepsilon}e^{-ip\cdot x}, \quad (156)$$

where $S(x)$ is the free quark propagator. As is easily seen in the above equation, $S(x)$ is singular for $x \to 0$. Since on the right-hand of Eq. (155) we have two $S(x)$'s in the first term, one $S(x)$ in the second and third, and none in the fourth, we realize that each term on the right-hand side of Eq. (155) is arranged in the order of decreasing singularity for $x \sim 0$. Equation (155) is clearly an example of the operator-product expansion (151).

The free quark propagator $S(x)$ is related to the free neutral scalar propagator $\Delta(x)$ defined previously in Eq. (143), i.e.,

$$S(x) = (i\not{\partial} + m)\Delta(x). \quad (157)$$

The explicit form of $\Delta(x)$ is given by Eq. (144). The leading singularity of $\Delta(x)$ may be extracted from Eq. (144) and is found to be independent of the quark mass,

$$\Delta(x) = \frac{1}{4\pi^2 i}\frac{1}{x^2 - i\varepsilon} + \text{less singular terms.} \tag{158}$$

The singularity lies on the light cone $x^2 \to 0$ and so is called a light-cone singularity. It is important to note here that the more singular the behavior of $\Delta(x)$ near the light cone, the larger the power of q^2 in the Fourier transform *sim* $\Delta(q)$ of $\Delta(x)$. The following formula is a typical example of this property in the one-dimensional Fourier transformation:

$$\int_{-\infty}^{\infty} dx \frac{e^{iqx}}{(x - i\varepsilon)^\alpha} = \frac{2\pi e^{i\alpha\pi/2}}{\Gamma(\alpha)}\theta(q)q^{\alpha-1}. \tag{159}$$

Hence it is enough for us to examine the most singular part of the c-number coefficients in Eq. (155) in order to see the dominant contribution of the current product to the matrix element of the physical reaction.

We extract the most singular part of Eq. (155) near the light cone by using Eqs. (157) and (158),

$$\begin{aligned} T[j_\mu(x)j_\nu(0)] = {} & \frac{x^2 g_{\mu\nu} - 2x_\mu x_\nu}{\pi^4(x^2 - i\varepsilon)^4} \\ & + \frac{ix^\lambda}{2\pi^2(x^2 - i\varepsilon)^2}\sigma_{\mu\lambda\nu\rho}O_V^\rho(x,0) \\ & + \frac{x^\lambda}{2\pi^2(x^2 - i\varepsilon)^2}\varepsilon_{\mu\lambda\nu\rho}O_A^\rho(x,0) \\ & + O_{\mu\nu}(x,0), \end{aligned} \tag{160}$$

where $O_V^\rho(x,0)$, $O_A^\rho(x,0)$, and $O_{\mu\nu}(x,0)$ are regular bilocal operators defined by

$$O_V^\mu(x,y) = :\bar\psi(x)\gamma^\mu\psi(y) - \bar\psi(y)\gamma^\mu\psi(x):, \tag{161}$$

$$O_A^\mu(x,y) = :\bar\psi(x)\gamma^\mu\gamma_5\psi(y) - \bar\psi(y)\gamma^\mu\gamma_5\psi(x):, \tag{162}$$

$$O_{\mu\nu}(x,y) = :\bar\psi(x)\gamma_\mu\psi(x)\bar\psi(y)\gamma_\nu\psi(y):, \tag{163}$$

and $\sigma_{\mu\lambda\nu\rho}$ is given by

$$\sigma_{\mu\lambda\nu\rho} = g_{\mu\lambda}g_{\nu\rho} + g_{\mu\rho}g_{\nu\lambda} - g_{\mu\nu}g_{\lambda\rho}. \tag{164}$$

It is important to note here that the operator-product expansion (160) provides us with a clear separation of the short-distance effects from the long-distance effects. In fact the singular c-number coefficients in the expansion characterize the short-distance behavior of the product of currents, while the regular bilocal operators include full information on the long-distance properties of the theory and are unimportant in the short-distance region.

We may transform Eq. (160) into a formula for the current commutator $[j_\mu(x), j_\nu(0)]$. For this purpose we first note that

$$T[j_\mu(x)j_\nu(0)] - T[j_\mu(x)j_\nu(0)]^\dagger = \varepsilon(x_0)[j_\mu(x), j_\nu(0)], \tag{165}$$

where we took into account that current j_μ is Hermitian, and $\varepsilon(x_0)$ is the sign function,

$$\varepsilon(x_0) = \frac{x_0}{|x_0|}. \tag{166}$$

We then use the fundamental relation

$$\frac{1}{x^2 - i\varepsilon} = \frac{P}{x^2} + i\pi\delta(x^2), \tag{167}$$

where P denotes the principal part. Differentiating Eq. (167) $n-1$ times with respect to x^2, we have

$$\frac{1}{(x^2 - i\varepsilon)^n} = \frac{P}{(x^2)^n} + i\pi\frac{(-1)^{n-1}}{(n-1)!}\delta^{(n-1)}(x^2), \tag{168}$$

where

$$\delta^{(n)}(x^2) = \frac{d^n}{d(x^2)^n}\delta(x^2). \tag{169}$$

From Eq. (168) we immediately obtain

$$\frac{1}{(x^2 - i\varepsilon)^n} - \frac{1}{(x^2 + i\varepsilon)^n} = 2\pi i\frac{(-1)^{n-1}}{(n-1)!}\delta^{(n-1)}(x^2). \tag{170}$$

Using Eqs. (160), (165), and (170), we finally obtain the desired formula,

$$\begin{aligned} \varepsilon(x_0)[j_\mu(x), j_\nu(0)] = {} & \frac{i}{3\pi^3}\big(2x_\mu x_\nu - x^2 g_{\mu\nu}\big)\delta^{(3)}(x^2) \\ & + \frac{1}{\pi}x^\lambda\delta^{(1)}(x^2)\sigma_{\mu\lambda\nu\rho}O_V^\rho(x,0) \\ & - \frac{i}{\pi}x^\lambda\delta^{(1)}(x^2)\varepsilon_{\mu\lambda\nu\rho}O_A^\rho(x,0) \\ & + O_{\mu\nu}(x,0) - O_{\nu\mu}(0,x). \end{aligned} \tag{171}$$

Now we apply the operator-product expansion to physical processes. Let us start with the application of Eq. (171) to the e^+e^- annihilation cross section. Using the standard method, the total cross section of e^+e^- annihilation can be expressed as

$$\sigma = \frac{e^4}{2s^3}l^{\mu\nu}w_{\mu\nu}, \tag{172}$$

where

$$l^{\mu\nu} = p_1^\mu p_2^\nu + p_1^\nu p_2^\mu - (q^2/2)g^{\mu\nu}, \tag{173}$$

$$w_{\mu\nu} = \int d^4x\, e^{iq\cdot x}\langle 0|[j_\mu(x), j_\nu(0)]|0\rangle, \tag{174}$$

Here p_1 (p_2) is the momentum of the electron (positron), and $q = p_1 + p_2$. We insert Eq. (171) in Eq. (191). Since O_V, O_A, and $O_{\mu\nu}$ are of the form of a normal product, we realize that only the first term on the right-hand side of Eq. (171) contributes to $w_{\mu\nu}$. Hence we have

$$w_{\mu\nu} = \frac{i}{3\pi^3}\left(g_{\mu\nu}\frac{\partial}{\partial q}\cdot\frac{\partial}{\partial q} - 2\frac{\partial}{\partial q^\mu}\cdot\frac{\partial}{\partial q^\nu}\right)I_3, \quad (175)$$

where nonleading contributions are neglected and I_n is given by

$$I_n = \int d^4x\, e^{iq\cdot x}\varepsilon(x_0)\delta^{(n)}(x^2). \quad (176)$$

After some calculation, I_n is found to be

$$I_n = \frac{i\pi^2}{4^{n-1}(n-1)!}(q^2)^{n-1}\varepsilon(q_0)\theta(q^2), \quad (177)$$

where $\theta(q^2)$ is the step function. We then obtain

$$w_{\mu\nu} = \frac{1}{6\pi}(q_\mu q_\nu - q^2 g_{\mu\nu})\varepsilon(q_0)\theta(q^2). \quad (178)$$

Substituting Eq. (178) into Eq. (186), we immediately find for the e^+e^- annihilation total cross section

$$\sigma = \frac{4\pi\alpha^2}{3s}. \quad (179)$$

In the above argument we started from the electromagnetic current given by Eq. (154) and so the quark was assumed to have unit charge, $Q = 1$. If we had started form the electromagneic current of the form

$$j_\mu(x) = \sum_{i=1}^{N_\mathrm{f}} Q_i \sum_{j=1}^{N_\mathrm{c}} :\bar{\psi}_{ij}(x)\gamma_\mu\psi_{ij}(x):, \quad (180)$$

with N_f quark flavors and N_c colors, we would have obtained

$$\sigma = \frac{4\pi\alpha^2}{3s}N_\mathrm{c}\sum_{i=1}^{N_\mathrm{f}} Q_i^2. \quad (181)$$

The above result is just the same as what is obtained from the parton model. Thus the free quark operator-product expansion (171) at short distances is essentially equivalent to the parton model. This is in a sense quite reasonable because the first term of the right-hand side of (171) comes from the first term of Eq. (155), which corresponds to the Feynman diagram depicted in Fig. 2. Taking the imaginary part of the Feynman amplitude corresponding to Fig. 2, we obtain the lowest order contribution of the electromagnetic interaction to the e^+e^- annihilation total cross section. Hence we have the parton model prediction of the cross section.

IV. PHYSICAL APPLICATIONS

Equipped with the operator-product expansion and renormalization group method we are ready to apply perturbative quantum chromodynamics to physical processes dominated by short-distance effects. As typical examples we deal with the total cross section of e^+e^- annihilations, deep inelastic structure functions, and hadron jet distributions in e^+e^- annihilations.

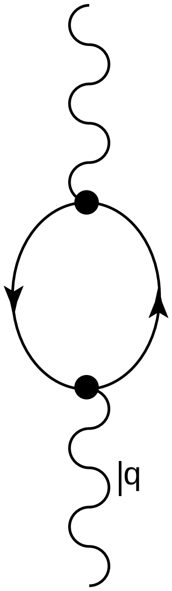


FIGURE 2

A. e^+e^- Annihilations

One of the simplest examples of the application of perturbative QCD is found in determining total cross sections for e^+e^- annihilation processes. Let us consider the process in which a positron e^+ and an electron e^- annihilate through the electromagnetic interaction, producing a number of hadrons. Here, for simplicity, we do not take into account the weak interaction effect which may become significant in the energy region of the weak neutral boson Z^0.

The process of e^+e^- annihilation is written as

$$e^+ + e^- \to X, \quad (182)$$

where X represents the final hadron system. The corresponding Feynman amplitude is given by

$$\langle X|T|e^+e^-\rangle = \bar{v}_{\lambda_2}(p_2)e\gamma^\mu u_{\lambda_1}(p_1)\frac{1}{q^2}\langle X|(-e)j_\mu(0)|0\rangle, \quad (183)$$

where $p_1(p_2)$ and $\lambda_1(\lambda_2)$ are the momentum and spin component of the electron (positron), respectively, with q the total momentum, $q = p_1 + p_2$; $u_\lambda(p)$ $[\bar{v}_\lambda(p)]$ is the Dirac spinor of the electron (positron) and $j_\mu(x)$ is the quark part of the electromagnetic current.

The total cross section for the annihilation process (182) can be written down as

$$\sigma = \frac{1}{2s}\frac{1}{4}\sum_{\lambda_1,\lambda_2}\sum_X (2\pi)^4\delta^4(p_X - q)|\langle X|T|e^+e^-\rangle|^2, \tag{184}$$

where the electron mass is neglected and

$$s = q^2 = (q_1 + q_2)^2. \tag{185}$$

Inserting Eq. (183) in Eq. (184), we ontain

$$\sigma = \frac{e^4}{2s^3} l^{\mu\nu} w_{\mu\nu}, \tag{186}$$

where

$$l^{\mu\nu} = p_1^\mu p_2^\nu + p_1^\nu p_2^\mu - (q^2/2)g^{\mu\nu}, \tag{187}$$

$$w_{\mu\nu} = \sum_X (2\pi)^4\delta^4(p_X - q)\langle 0|j_\mu(0)|X\rangle\langle X|j_\nu(0)|0\rangle. \tag{188}$$

It is easily shown that $w_{\mu\nu}$ can be rewritten as

$$w_{\mu\nu} = \int d^4x e^{iq\cdot x}\langle 0|j_\mu(x)j_\nu(0)|0\rangle. \tag{189}$$

For the process $e^+ + e^- \to X$ to be physical, the total energy q_0 of the initial state should be positive, and then we can show

$$\int d^4x e^{iq\cdot x}\langle 0|j_\nu(0)j_\mu(x)|0\rangle = 0. \tag{190}$$

With Eq. (190), Eq. (189) can be written in the form

$$w_{\mu\nu} = \int d^4x e^{iq\cdot x}\langle 0|[j_\mu(x)j_\nu(0)]|0\rangle. \tag{191}$$

Thus the total cross section for e^+e^- annihilation is expressed in terms of the current commutator. In the center-of-mass system, we have $q = (q_0, 0, 0, 0)$. Bearing in mind high-energy annihilations, we let $q_0 \to \infty$ and we find that only the region $x_0 \sim 0$ makes a major contribution to the integral (191) according to the Riemann–Lebesgue theorem. On the other hand, the integral (191) has a support only when $x^2 \geq 0$ due to the causality requirement, so that $x_0 \sim 0$ implies $x \sim 0$. Hence we conclude that the total cross section for high-energy e^+e^- annihilations is governed by the current commutator at short distances.

For later convenience we further rewrite Eq. (186). The general tensor structure of $w_{\mu\nu}$ may be easily deduced following the requirements of Lorentz invariance and current conservation. We find that $w_{\mu\nu}$ is expressed in terms of only one invariant amplitude $w(q^2)$,

$$w_{\mu\nu} = \left(q_\mu q_\nu - q^2 g_{\mu\nu}\right)\frac{1}{6\pi}w(q^2), \tag{192}$$

where the extra factor $1/(6\pi)$ is attached for later convenience. Substituting Eq. (192) into Eq. (186), we have

$$\sigma = \frac{4\pi\alpha^2}{3s}w(s). \tag{193}$$

where $\alpha = e^2/(4\pi)$. On the other hand, it is easy to show that the total cross section for the process

$$e^+ + e^- \to \mu^+ + \mu^- \tag{194}$$

in the lowest order of the electromagnetic interaction is equal to

$$\sigma_{\mu\nu} = \frac{4\pi\alpha^2}{3s}, \tag{195}$$

where the electron and muon masses are neglected. It is customary to define so-called the R ratio

$$R = \frac{\sigma}{\sigma_{\mu\mu}} = w(s), \tag{196}$$

to discuss the high-energy e^+e^- annihilation process. The R ratio is closely related to the current commutator at short distances. To show, this we use Eqs. (191), (192) and (196) to reexpress the R ratio as

$$R = -\frac{2\pi}{q^2}\int d^4x e^{iq\cdot x}\big\langle 0\big|j_\mu(x)j^\mu(0)\big|0\big\rangle. \tag{197}$$

We consider the e^+e^- annihilation at very high center-of-mass energies (large $\sqrt{q^2}$) so that all the relevant quark masses are negligible compared with $\sqrt{q^2}$. Then R is a function only of the center-of-mass energy squared, $s = q^2$, the renormalized coupling constant g, and the renormalization scale μ,

$$R = R(s/\mu^2, g). \tag{198}$$

Here the dependence of R on s and μ is given by the ratio s/μ^2 for dimensional reasons.

The operator j_μ in Eq. (197) is the electromagnetic current, which is conserved, and hence its anomalous dimension vanishes. Accordingly, the renormalization group equation for the R ratio reads

$$\left[\mu\frac{\partial}{\partial\mu} + \beta(g)\frac{\partial}{\partial g}\right]R\left(\frac{s}{\mu^2}, g\right) = 0. \tag{199}$$

The general solution of Eq. (199) is easily found to be

$$R\left(\frac{s}{\mu^2}, g\right) = R(1, \bar{g}(s)), \tag{200}$$

where the running coupling constant $\bar{g}(s)$ is defined in terms of the β-function such that

$$\frac{\partial\bar{g}}{\partial t} = \beta(\bar{g}), \qquad \bar{g}(\mu^2) = g, \tag{201}$$

with $t = (1/2)\ln(s/\mu^2)$.

The meaning of Eq. (200) is obvious: the explicit s dependence of the R ratio computed by using the coupling g can be completely absorbed into the s dependence of the running coupling constant $\bar{g}(s)$. In asymptotically free field theories, the running coupling constant $\bar{g}(s)$ for large s is found to be small; thus the validity of the perturbative calculation of R is guaranteed.

The R ratio expressed in terms of the coupling constant g renormalized at scale μ contains, in general, large logs, $\ln(s/\mu^2)$, for large s in each term of the perturbative expansion and hence the effectiveness of the perturbative calculation is spoiled. According to Eq. (200), however, the calculation is drasticaly improved if we employ the coupling constant renormalized at the scale of the relevant energy $\sqrt{s}$.

Let us look into the details of the above statement. The R ratio $R(s/\mu^2, g)$ is given by the perturbative calculation in the form

$$R\left(\frac{s}{\mu^2}, g\right) = \sum_i Q_i^2[1+a(s/\mu^2)g^2+b(s/\mu^2)g^4+\cdots], \tag{202}$$

where the index i runs over colors and flavors of quarks. The coefficients a, b, ... in general include large logs, $\ln(s/\mu^2)$. Equation (200) indicates that, if g is replaced by $\bar{g}(s)$, these large logs in the coefficient disappear, i.e.,

$$R\left(\frac{s}{\mu^2}, g\right) = \sum_i Q_i^2[1+a(1)\bar{g}(s)^2+b(1)\bar{g}(s)^4+\cdots]. \tag{203}$$

The expression (203) is much better than Eq. (202) in two respects: its expansion coefficients are smaller than those in Eq. (202) and the expansion parameter $\bar{g}(s)$ is smaller than g for large s ($s \gg \mu^2$) according to the property of asymptotic freedom.

We shall show how to calculate $a(1)$ in Eq. (203) in perturbative QCD. The strategy for computing $a(1)$ is first to calculate the R ratio to order g^2 by using the coupling constant g renormalized at the scale μ and then set $\mu^2 = s$ to obtain $a(1)$.

The Feynman diagrams contributing to the total cross section of the e^+e^- annihilation are shown in Fig. 3. The total cross section σ receives separate contributions from the final states $q\bar{q}, q\bar{q}G, q\bar{q}q\bar{q}, q\bar{q}GG, \ldots$, with $q, \bar{q}$, and G denoting the quark, antiquark, and gluon, respectively. The contribution up to order g^2 may be represented as in Fig. 4 and is split into three parts: the Born cross section σ_B (Fig. 4a), the virtual (one-loop) gluon contribution σ_V (Fig. 4b, c), and the real-gluon-emission cross section σ_R (Fig. 4d). Denoting the full cross section to order g^2 by σ, we have

$$\begin{aligned}\sigma &= Z_2^2\sigma_B + \sigma_V + \sigma_R \\ &= \sigma_B + \tilde{\sigma}_V + \sigma_R,\end{aligned} \tag{204}$$

where $\tilde{\sigma}_V = \sigma_V + (Z_2^2 - 1)\sigma_B$ with Z_2 the field renormalization constant associated with the quark extrnal lines. The factor Z_2^2 in Eq. (204) is necessary since the field renormalization constant $Z_2^{1/2}$ for each quark external line should

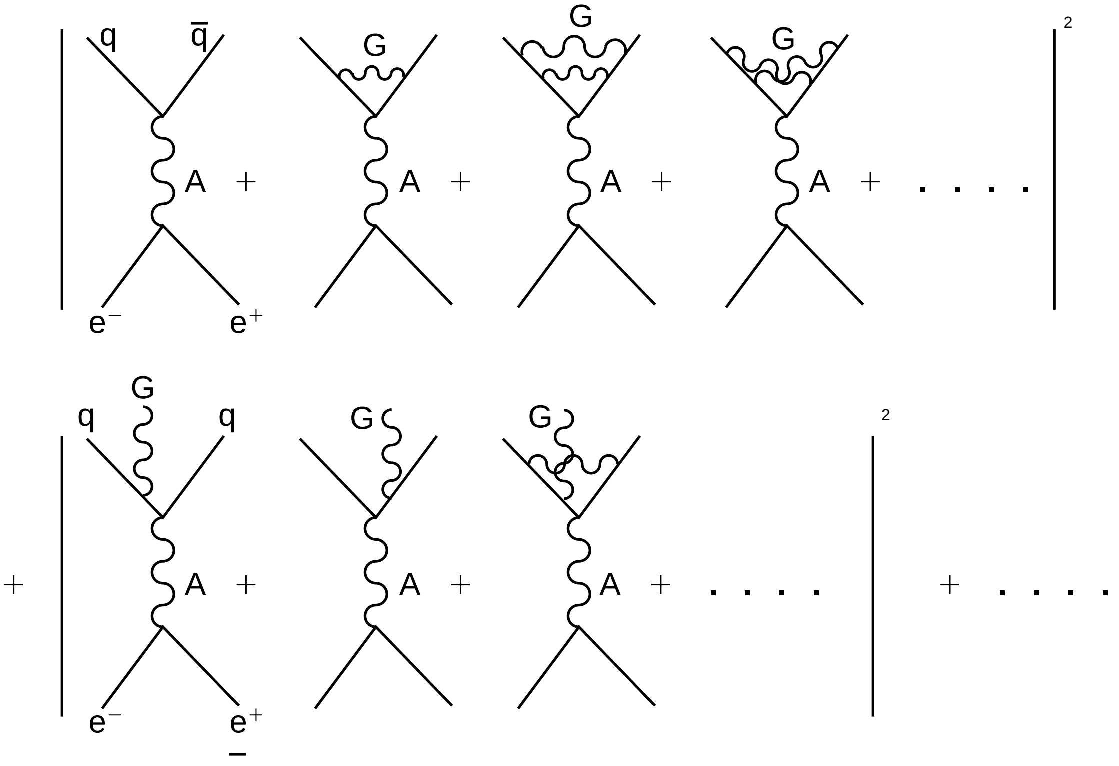


FIGURE 3

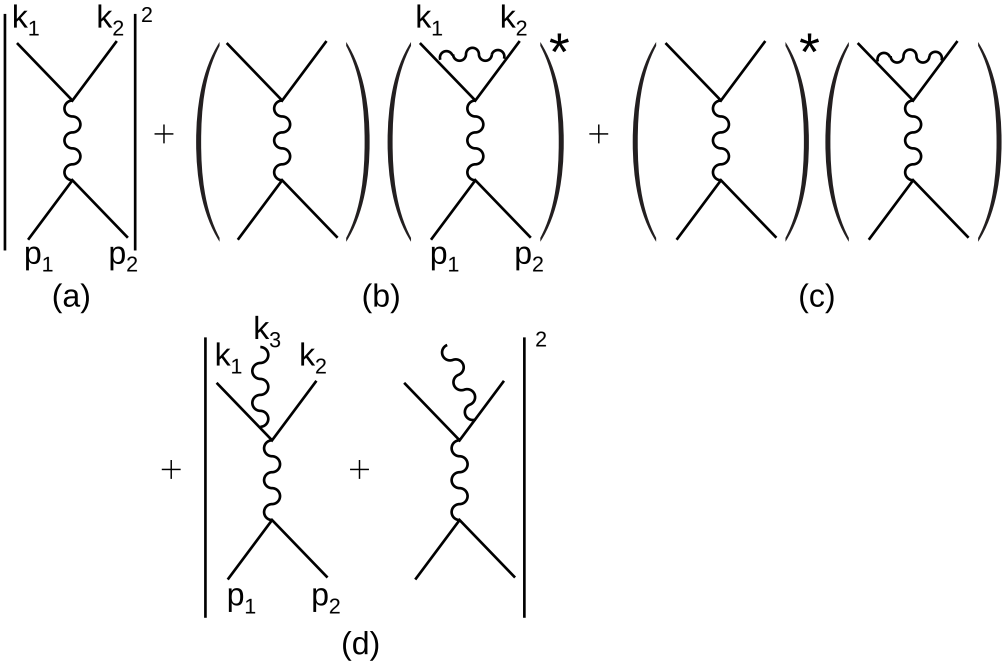


FIGURE 4

be included in the expression of the renormalized S-matrix elements as a renormalized truncated Green function.

The Born cross section $\sigma_{\rm B}$ can be obtained by neglecting electron and quark messes,

$$\sigma_{\rm B} = \frac{4\pi\alpha^2}{3s}\sum_i Q_i^2. \tag{205}$$

The one-loop contribution $\sigma_{\rm V}$ to the $e^+e^- \to q\bar{q}$ cross section of Figs. 4b and 4c is calculated in the following way. We consider the e^+e^- annihilations at very high energies so that quark masses are practically negligible. In the following calculations all the quarks are regarded as massless. The contribution of Figs. 4b and 4c can be written in such a way that

$$\sigma_{\rm V} = \frac{1}{8s}\int \frac{d^3k_1}{(2\pi)^3 2k_{10}}\frac{d^3k_2}{(2\pi)^3 2k_{20}}(2\pi)^4\delta^4 \times (k_1 + k_2 - p_1 - p_2)F_{\rm V}, \tag{206}$$

where $F_{\rm V}$ is given by

$$F_{\rm V} = \left(\sum_i Q_i^2\right)\frac{e^4}{q^4}{\rm Tr}\left[\not{p}_2\gamma^\mu\not{p}_1\gamma^\nu\right]{\rm Tr}[\not{k}_1\Lambda_\mu\not{k}_2\gamma_\nu] + {\rm c.c.}, \tag{207}$$

with c.c. representing the complex conjugate of the first term and Λ_μ the one-loop vertex part corresponding to Fig. 5,

$$\Lambda_\mu = g^2 C_{\rm F}\int\frac{d^Dk}{(2\pi)^D i}\frac{1}{k^2}\gamma_\lambda\frac{1}{\not{k}-\not{k}_1}\gamma_\mu\frac{1}{\not{k}+\not{k}_2}\gamma^\lambda. \tag{208}$$

Note also that we use the Feynman gauge in the present calculation, and we calculate Z_2 on the mass shell of quarks where quarks are massless. Here naturally we meet with the infrared divergence (mass singularity) arising from the vanishing quark mass. We shall regularize the mass singularity by means of dimensional regularization. For $p^2 = 0$ the quark self-energy part $\Sigma(p)$ in the Feynman gauge reads

$$\Sigma(p)|_{p^2=0} = g^2 C_{\rm F}(D-2)\not{p}\int_0^1 dx(1-x)\int\frac{d^Dk}{(2\pi)^D i}\frac{1}{k^4}$$
$$= \frac{g^2}{(4\pi)^2}C_{\rm F}\not{p}\left(\frac{1}{\varepsilon}-\frac{1}{\varepsilon'}\right), \tag{209}$$

where ε' and ε are equal to the parameter $(4-D)/2$ and serve to regularize the ultraviolet and infrared divergences,

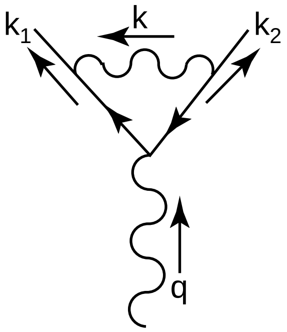


FIGURE 5

respectively. From Eq. (209) we obtain the renormalization constant Z_2 to one-loop order,

$$Z_2 = 1 + \frac{g^2}{(4\pi)^2} C_{\mathrm{F}} \left(\frac{1}{\varepsilon} - \frac{1}{\varepsilon'} \right). \tag{210}$$

The integration in Eq. (208) is performed similarly and results in

$$\begin{aligned} \Lambda_\mu = {} & \gamma_\mu \frac{g^2}{8\pi^2} C_{\mathrm{F}} \left(\frac{4\pi\mu^2}{-q^2} \right)^{\varepsilon} \Gamma(1+\varepsilon) B(1-\varepsilon, 2-\varepsilon) \\ & \times \left(\frac{1}{\varepsilon'} - \frac{2}{\varepsilon^2} - 2 \right), \end{aligned} \tag{211}$$

where μ is the mass scale introduced to make the coupling g dimensionless and

$$q = p_1 + p_2. \tag{212}$$

As expected, the ultraviolet divergences present in Eq. (211) cancel out in $\tilde{\sigma}_{\mathrm{V}}$ on account of Eq. (210). Inserting Eq. (211) into Eq. (207) and taking account of Eqs. (205) and Eq. (210), we find

$$\tilde{\sigma}_{\mathrm{V}} = A_{\mathrm{V}} \sigma_{\mathrm{B}}, \tag{213}$$

where A_{V} is given by

$$\begin{aligned} A_{\mathrm{V}} = {} & \frac{\alpha_s}{\pi} C_{\mathrm{F}} \left(\frac{4\pi\mu^2}{s} \right)^{\varepsilon} \frac{\cos \pi\varepsilon}{\Gamma(1-\varepsilon)} \\ & \times \left(-\frac{1}{\varepsilon^2} - \frac{3}{2\varepsilon} - 4 + O(\varepsilon) \right), \end{aligned} \tag{214}$$

where α_s is the QCD coupling constant defined by $\alpha_s = g^2/(4\pi)$.

The calculation of σ_{R} goes as follows. The cross section σ_{R} corresponding to Fig. 4d is written in the form

$$\begin{aligned} \sigma_{\mathrm{R}} = {} & \frac{1}{8s} \int \prod_{i=1}^{3} \frac{d^{D-1}k_i}{(2\pi)^{D-1} 2k_{i0}} (2\pi)^D \delta^D \\ & \times \left(\sum_{i=1}^{3} k_i - p_1 - p_2 \right) F_{\mathrm{R}}, \end{aligned} \tag{215}$$

where

$$F_R = -\left(\sum_i Q_i^2 \right) \frac{e^4}{q^4} g^2 C_{\mathrm{F}} \operatorname{Tr}\left[\not{p}_2 \gamma^\mu \not{p}_1 \gamma^\nu \right] \operatorname{Tr}\left[\not{k}_1 S_\lambda \not{k}_2 S_\nu^\lambda \right], \tag{216}$$

$$S_{\mu\nu} = \gamma_\mu \frac{-1}{\not{k}_1 + \not{k}_3} \gamma_\nu + \gamma_\nu \frac{1}{\not{k}_2 + \not{k}_3} \gamma_\mu. \tag{217}$$

In Eq. (215) we worked in D dimensions ratther than in four dimensions because we anticipate possible infrared divergences. We introduce tensors $G_{\mu\nu}$, $L^{\mu\nu}$, and $I_{\mu\nu}$,

$$G_{\mu\nu} = \operatorname{Tr}\left[\not{k}_1 S_\lambda \not{k}_2 S_\nu^\lambda \right], \tag{218}$$

$$L^{\mu\nu} = \operatorname{Tr}\left[\not{p}_2 \gamma^\mu \not{p}_1 \gamma^\nu \right] = 4\left(p_1^\mu p_2^\nu + p_1^\nu p_2^\mu - \frac{q^2}{2} g^{\mu\nu} \right), \tag{219}$$

$$I_{\mu\nu} = \int \prod_{i=1}^{3} \frac{d^{D-1}k_i}{2k_{i0}} \delta^D \left(\sum_{i=1}^{3} k_i - q \right) G_{\mu\nu}. \tag{220}$$

The cross section σ_{R} is then expressed in the form

$$\sigma_{\mathrm{R}} = \frac{-e^4 g^2}{8s(2\pi)^{2D-3} q^4} \left(\sum_i Q_i^2 \right) C_{\mathrm{F}} L^{\mu\nu} I_{\mu\nu}. \tag{221}$$

As can be seen in Eq. (220), $I_{\mu\nu}$ depends only on q_μ and satisfies the following condition corresponding to the conservation of the electromagnetic current:

$$q^\mu I_{\mu\nu} = 0. \tag{222}$$

Hence the general form of $I_{\mu\nu}$ is given by

$$I_{\mu\nu} = I(q^2) \left(\frac{q_\mu q_\nu}{q^2} - g_{\mu\nu} \right), \tag{223}$$

with $I(q^2) = -g^{\mu\nu} I_{\mu\nu}/(D-1)$. By means of Eq. (223) we obtain

$$L^{\mu\nu} I_{\mu\nu} = \frac{D-2}{D-1} q^2 g^{\mu\nu} I_{\mu\nu}. \tag{224}$$

On the other hand, we find after some calculation

$$g^{\mu\nu} G_{\mu\nu} = -8(1-\varepsilon) \frac{x_1^2 + x_2^2 - \varepsilon x_3^2}{(1-x_1)(1-x_2)}, \tag{225}$$

where $\varepsilon = (4-D)/2$ and

$$x_i = 2k_i \cdot q/q^2 \qquad (i = 1, 2, 3). \tag{226}$$

The three-body phase volume for $g^{\mu\nu} I_{\mu\nu}$ in D dimensions may be rewritten in terms of the variables x_i so that

$$\begin{aligned} g^{\mu\nu} I_{\mu\nu} = {} & \frac{\pi(\pi s)^{1-2\varepsilon}}{4\Gamma(2-2\varepsilon)} \int_0^1 \prod_{i=1}^{3} (1-x_i)^{-\varepsilon} \, dx_i \, \delta \\ & \times \left(2 - \sum_{i=1}^{3} x_i \right) g^{\mu\nu} G_{\mu\nu}. \end{aligned} \tag{227}$$

Combining Eqs. (221), (224), (225), and (227), we arrive at

$$\begin{aligned} \sigma_{\mathrm{R}} = {} & \left(\sum_i Q_i^2 \right) \alpha^2 \alpha_s C_{\mathrm{F}} \frac{2}{s} \left(\frac{4\pi\mu}{s} \right)^{2\varepsilon} \\ & \times \frac{(1-\varepsilon)^2}{(3-2\varepsilon)\Gamma(2-2\varepsilon)} K. \end{aligned} \tag{228}$$

where μ comes from the mass dimension of the coupling constant g in D dimensions and

$$K = \int_0^1 \prod_{i=1}^{3}(1-x_i)^{-\varepsilon}\, dx_i\, \delta\left(2-\sum_{i=1}^{3} x_i\right) \frac{x_1^2+x_2^2-\varepsilon x_3^2}{(1-x_1)(1-x_2)}. \tag{229}$$

The constant K can be calculated analytically to order ε^0, i.e.,

$$K = \left(\frac{4}{\varepsilon^2} - \frac{12}{\varepsilon} + 10 - 4\varepsilon\right) B(1-\varepsilon, 2-2\varepsilon) B \times (1-\varepsilon, 1-\varepsilon) + O(\varepsilon). \tag{230}$$

We wish to express Eq. (228) in the form of

$$\sigma_{\mathrm{R}} = A_{\mathrm{R}} \sigma_{\mathrm{B}}, \tag{231}$$

with A_{R} to be determined. For this purpose we need to find the expression for the Born cross section σ_{B} in D dimensions. We repeat the calculation of σ_{B} in D-dimensional space-time and obtain

$$\sigma_{\mathrm{B}} = \frac{4\pi\alpha^2}{3s}\left(\sum_i Q_i^2\right)\left(\frac{4\pi}{s}\right)^{\varepsilon} \frac{3(1-\varepsilon)\Gamma(2-\varepsilon)}{(3-2\varepsilon)\Gamma(2-2\varepsilon)}. \tag{232}$$

Comparing Eq. (228) with (232), we find for A_{R} of Eq. (231)

$$A_{\mathrm{R}} = \frac{\alpha_s}{\pi} C_{\mathrm{F}} \left(\frac{4\pi\mu^2}{s}\right)^{\varepsilon} \frac{\cos\pi\varepsilon}{\Gamma(1-\varepsilon)} \times \left(\frac{1}{\varepsilon^2} + \frac{3}{2\varepsilon} + \frac{19}{4} + O(\varepsilon)\right). \tag{233}$$

Substituting Eqs. (213) and (231) into Eq. (204), we finally obtain

$$\sigma = (1 + A_{\mathrm{V}} + A_{\mathrm{R}})\sigma_{\mathrm{B}} = \left(1 + \frac{3}{4} C_{\mathrm{F}} \frac{\alpha_s}{\pi}\right)\sigma_{\mathrm{B}}. \tag{234}$$

In this result we clearly see that the infrared divergences present in A_{V} and A_{R} just cancel out leaving a finite one-loop effect. We thus finally obtain $a(s/\mu^2)$, which was defined in Eq. (202).

It should be noted here that up to this order, $a(s/\mu^2)$ is independent of large logs, $\ln(s/\mu^2)$, and so

$$a(s/\mu^2) = a(1). \tag{235}$$

Under this circumstance we may, according to Eq. (200), simply replace α_s in Eq. (234) by $\bar{\alpha}_s$, the running coupling constant,

$$\bar{\alpha}_s = \frac{\bar{g}^2}{\pi}, \tag{236}$$

and obtain

$$R\left(\frac{s}{\mu^2}, g\right) = \sum_i Q_i^2 \left(1 + \frac{3}{4} C_{\mathrm{F}} \frac{\bar{\alpha}_s}{\pi}\right). \tag{237}$$

Owing to asymptotic freedom, the running coupling constant of QCD decreases logarithmically as the relevant mass scale grows. Accordingly, Eq. (237) tells us that the R ratio in QCD approaches the parton-model prediction as $s \to \infty$.

B. $q\bar{q}$ Jets in e^+e^- Annihilations

We shall show in the present section that hadronic jet phenomena are dominated by short-distance effects, so that perturbative QCD may be safely applied to the discussion of jet processes. For this purpose we present the proof of the cancellation of the infrared divergences in the jet cross sections since the infrared divergences reflect the long-distance nature of QCD.

Here we confine ourselves to hadronic jets arising from the quark–antiquark pair production in e^+e^- annihilations. In order that the hadronic jets flow from the quark–antiquark pair, the quark and antiquark are required not to lose too much energy in their direction by the emission of gluons and quark pairs. In other words it is necessary to show that most of the annihilation energy is deposited along the direction of the quark and antiquark. The first step of the argument is to give a precise definition of the jet cross section and then to prove that the dangerous infrared (soft and collinear) divergences are controllable in size. The jet generated by the above QCD mechanism is often called the Sterman–Weinberg jet or the QCD jet.

We have calculated in Section IV.A the total cross section of e^+e^- annihilations to which the QCD diagrams shown in Fig. 3 contribute. The contribution up to order g^2 was represented in Fig. 4. Here we discuss the jet contibution up to the same order as above. The diagrams contibuting to jets are the same as in Fig. 4 though the kinematical region in the final state is restricted. We define the two-jet event in such a way that most of the available energy $\sqrt{s}$, i.e., $(1-\Delta)\sqrt{s}$ with $\Delta \ll 1$, is deposited on two cones of small half-angle δ (see Fig. 6). We consider the angular distribution of the final quarks, i.e., the differential cross section $d\sigma/d\Omega$ in the solid angle Ω specified by polar and azimuthal angles θ and ϕ, respectively. The Born contribution $(d\sigma/d\Omega)_{\mathrm{B}}$ corresponding to Fig. 4a is given by

$$\left(\frac{d\sigma}{d\Omega}\right)_{\mathrm{B}} = \frac{\alpha^2}{4s}\sum_i Q_i^2(1+\cos^2\theta). \tag{238}$$

The virtual (one-loop) contribution $(d\sigma/d\Omega)_{\mathrm{V}}$ corresponding to Figs. 4b and 4c is directly obtained in parallel with the previous argument in deriving Eq. (213),

$$\left(\frac{d\sigma}{d\Omega}\right)_{\mathrm{V}} = A_{\mathrm{V}}\left(\frac{d\sigma}{d\Omega}\right)_{\mathrm{B}}, \tag{239}$$

where A_{V} is given by Eq. (214).

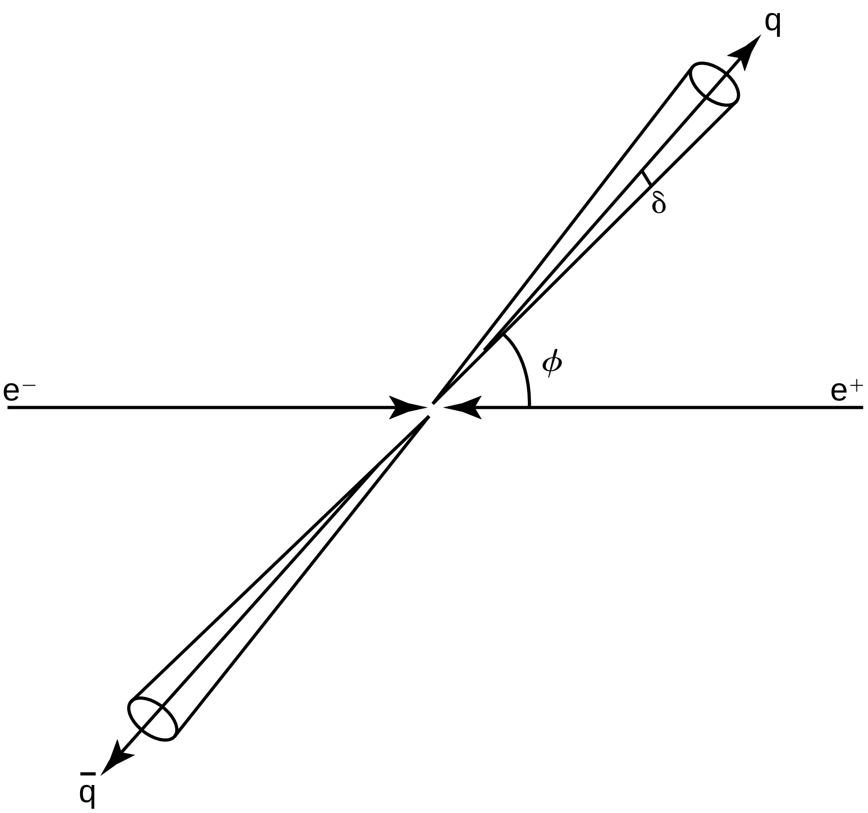


FIGURE 6

Now comes the real-gluon-emission contribution $(d\sigma/d\Omega)_{\rm R}$ corresponding to Fig. 4d, which occupies the major part of the argument in the rest of the present subsection. The differential cross section $(d\sigma/d\Omega)_{\rm R}$ is defined in a similar way as in Eq. (215) by

$$\int \left(\frac{d\sigma}{d\Omega}\right)_{\rm R} d\Omega = \frac{1}{8s}\int_R \prod_{i=1}^{3} \frac{d^{D-1}k_i}{(2\pi)^{D-1}2k_{i0}}(2\pi)^D \delta^D \times \left(\sum_{i=1}^{3} k_i - p_1 - p_2\right) F_{\rm R}, \qquad (240)$$

where $F_{\rm R}$ is given by Eq. (216) and the integral region R is specified by the following conditions: The emitted gluon is either soft (i.e., $x_3 \le \Delta$) or collinear to one of the quarks (i.e., $\theta_{13}, \theta_{23} < 2\delta$), where x_3 is defined by Eq. (226) and $\theta_{13}(\theta_{23})$ is the angle between the gluon and the quark (antiquark) as shown in Fig. 7. It should be noted here that the restriction of the phase space to R in Eq. (240) corresponds to taking the degenerate state of the quark and gluon, and the infrared divergences (soft and collinear) are known to cancel out in the following sum:

$$\frac{d\sigma}{d\Omega} = \left(\frac{d\sigma}{d\Omega}\right)_{\rm B} + \left(\frac{d\sigma}{d\Omega}\right)_{\rm V} + \left(\frac{d\sigma}{d\Omega}\right)_{\rm R}. \qquad (241)$$

It is convenient to define new variables ζ_1 and ζ_2 through

$$\zeta_1 = \frac{1}{2}(1-\cos\theta_{13}) = \frac{1-x_2}{x_1x_3}, \qquad (242)$$

$$\zeta_2 = \frac{1}{2}(1-\cos\theta_{23}) = \frac{1-x_1}{x_1x_3}, \qquad (243)$$

where use has been made of the relation

$$(k_1+k_3)^2 = \frac{1}{4}x_1x_3s(1-\cos\theta_{13}) = (q-k_2)^2 = s(1-x_2), \text{ etc.} \qquad (244)$$

Since $\theta_{13}, \theta_{23} \le 2\delta$ in R, we have

$$\zeta_1, \zeta_2 \le \sin^2\delta. \qquad (245)$$

Only two variables are independent among the five variables x_1, x_2, x_3, ζ_1, and ζ_2 and so we choose ζ_1 and x_3 as independent variables. Then the region R is specified by

$$\frac{1-\sin^2\delta}{1-x_3(2-x_3)\sin^2\delta} \le \zeta_1 \le \sin^2\delta, \qquad 0 \le x_3 \le \Delta, \qquad (246)$$

where the lower bound of ζ_1 comes from the condition $\zeta_2 < \sin^2\delta$. In Fig. 8 the kinematical region R given by Eq. (246) is shown in the ζ_1–x_3 plane. Neglecting terms of order δ^2, we express $(d\sigma/d\Omega)_{\rm R}$ in the form of an angular distribution with respect to the quark direction,

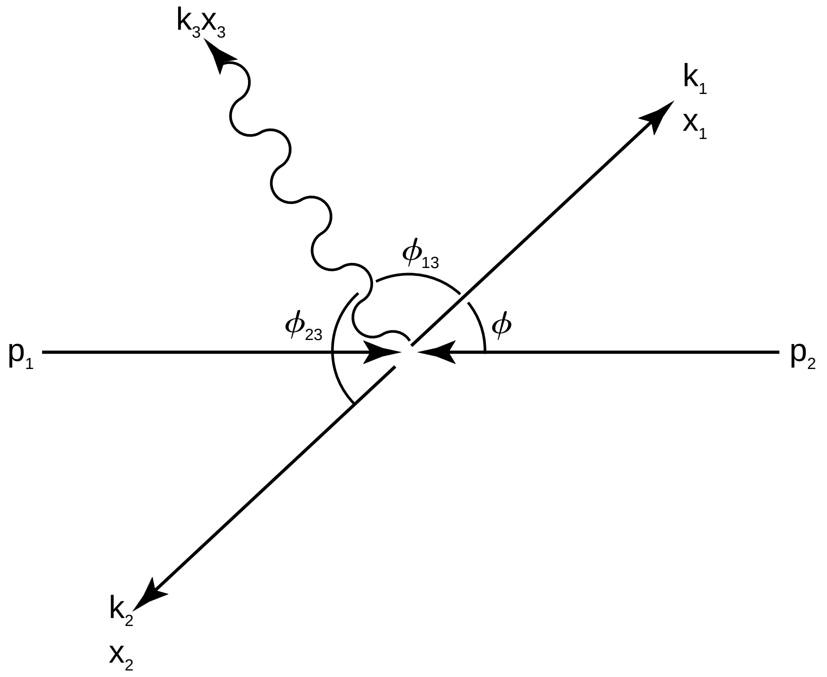


FIGURE 7

$$\left(\frac{d\sigma}{d\Omega}\right)_{\rm R} = \frac{3}{16\pi}(1+\cos^2\theta)\alpha^2\alpha_s C_{\rm F}\left(\sum_i Q_i^2\right)\frac{2}{s} \times \left(\frac{4\pi\mu}{s}\right)^{2\varepsilon}\frac{(1-\varepsilon)^2}{(3-2\varepsilon)\Gamma(2-2\varepsilon)}K_R, \quad (247)$$

where

$$K_R = \int_R \left[\prod_{i=1}^3 (1-x_i)^{-\varepsilon}\,dx_i\right]\delta\left(2-\sum_{i=1}^3 x_i\right)\rho, \quad (248)$$

$$\rho = \frac{x_1^2+x_2^2-\varepsilon x_3^2}{(1-x_1)(1-x_2)}. \quad (249)$$

Note that the slight deviation from the above angular distribution is expected if a precise calculation is made. Probably the easiest way of calculating $K_{\rm R}$ in Eq. (248) is the following: We first note that

$$K_R = K - K_{\bar R}, \quad (250)$$

where K is the quantity corresponding to K_R integrated over the whole phase space and is already given in Eq. (230) and $K_{\bar R}$ is obtained by integrating the integrand of Eq. (248) over the region $\bar R$ which is obtained by eliminating R from the whole phase space as shown in Fig. 8. Since there is no infrared singularity in the region $\bar R$, we may put $\varepsilon = 0$ in $K_{\bar R}$ and then the calculation turns out to be straightforward. Changing the variables to ζ_1 and x_3, we obtain

$$K_{\bar R} = \int_\Delta^1 dx_3 x_3(1-x_3)\int_{\zeta_{\rm min}}^{\zeta_{\rm max}} d\zeta_1 \frac{\rho}{(1-x_3\zeta_1)^2}, \quad (251)$$

where $\zeta_{\rm max}$ and $\zeta_{\rm min}$ are respectively the maximum and minimum values of ζ_1 given in Eq. (246), and ρ defined in Eq. (249) is rewritten in terms of ζ_1 and x_3 as

$$\rho = \frac{(1-x_3)^2+(1-x_3(2-x_3)\zeta_1)^2}{x_3^2(1-x_3)\zeta_1(1-\zeta_1)}. \quad (252)$$

After some calculation we find

$$K_{\bar R} = 2\left(4\ln\delta\ln\Delta + 3\ln\delta - \frac{7}{4}+\frac{\pi^2}{3}\right) + O(\delta,\Delta). \quad (253)$$

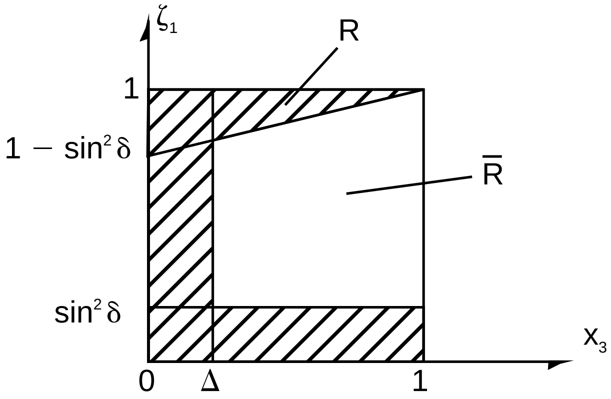


FIGURE 8

Hence we have

$$K_R = 4B(1-\varepsilon, 2-2\varepsilon)B(1-\varepsilon, 1-\varepsilon) \times \left(\frac{1}{\varepsilon^2} - \frac{3}{\varepsilon} + \frac{17}{4} - \frac{\pi^2}{3} - 4\ln\delta\ln\Delta - 3\ln\delta\right). \tag{254}$$

Exactly in the same way as in Eq. (232) we calculate the Born contribution to $d\sigma/d\Omega$ in D dimensions, which results in

$$\left(\frac{d\sigma}{d\Omega}\right)_{\mathrm{B}} = \frac{\alpha^2}{4s}\sum_i Q_i^2(1+\cos^2\theta)\left(\frac{4\pi}{s}\right)^{\varepsilon} \times \frac{3(1-\varepsilon)\Gamma(2-\varepsilon)}{(3-2\varepsilon)\Gamma(2-2\varepsilon)}. \tag{255}$$

Using Eqs. (247), (254), and (255) we finally obtain

$$\left(\frac{d\sigma}{d\Omega}\right)_{\mathrm{R}} = \left(\frac{d\sigma}{d\Omega}\right)_{\mathrm{B}} \frac{\alpha_s}{\pi} C_{\mathrm{F}} \left(\frac{4\pi\mu^2}{s}\right)^{\varepsilon} \frac{\cos\pi\varepsilon}{\Gamma(1-\varepsilon)} \left(\frac{1}{\varepsilon^2} + \frac{3}{2\varepsilon} + \frac{13}{2} - \frac{\pi^2}{3} - 4\ln\delta\ln\Delta - 3\ln\delta\right). \tag{256}$$

Combining Eqs. (238), (239), and (256), we see that the infrared divergences just cancel out, and find

$$\frac{d\sigma}{d\Omega} = \left(\frac{d\sigma}{d\Omega}\right)_{\mathrm{B}} \left[1 - \frac{\alpha_s}{\pi} C_{\mathrm{F}} \left(4\ln\delta\ln\Delta + 3\ln\delta + \frac{\pi^2}{3} - \frac{5}{2}\right)\right]. \tag{257}$$

According to Eq. (257), we realize that the order-α_s correction to the two jets from the quark pair is controllable in size within the framework of perturbation theory and the hadronic pair jets appproximately in the direction of the quark and antiquark. Thus the angular distribution of the hadronic pair jets reflects the spin-$1/2$ nature of the constituents.

A similar argument as above may be made for hadronic jets originating from gluon sources. It has been shown that the same infrared cancellation as in the quark jets takes place in the gluon jets. Hence the hadronic jet from the gluon should also be observed experimentally. In fact, clear signals of three jets from the quark, antiquark, and gluon have been observed in e^+e^- annihilation processes.

SEE ALSO THE FOLLOWING ARTICLES

GREEN'S FUNCTIONS • GROUP THEORY, APPLIED • PERTURBATION THEORY • QUANTUM MECHANICS

BIBLIOGRAPHY

Bromley, D. A., and Schäfer, A. (1994). "Quantum Chromodynamics," Springer-Verlag, Berlin.

Close, F. (1997). "The Cosmic Onion: Quarks and the Nature of the Universe," Springer-Verlag, Berlin.

Fernandez, F. M. (2001). "Introduction to Perturbation Theory in Quantum Mechanics," CRS Press, Boca Raton, FL.

Forshaw, J. R., and Ross, D. A. (1997). "Quantum Chromodynamics and the Pomeron," Cambridge University Press, Cambridge.

Henyey, F. (1994). "Quantum Chromodynamics," Springer-Verlag, Berlin.

Manohar, A. V., and Wise, M. B. (2000). "Heavy Quark Physics," Cambridge University Press, Cambridge.

Reinhardt, H., and Alkofer, R. (1995). "Chiral Quark Dynamics," Springer-Verlag, Berlin.

Quantum Hall Effect

J. K. Jain
Pennsylvania State University

GLOSSARY

Composite fermion (CF) The bound state of an electron and an even number ($2p$) of quantum mechanical vortices.

Filling factor (ν) The ratio of the number of electrons to the number of flux quanta (a flux quantum is defined as $\phi_0 = hc/e$) penetrating the sample. It is nominally equal to the number of filled Landau levels.

Hall effect Generation of a voltage transverse to the direction of current flow in the presence of a magnetic field. The ratio of the transverse voltage to the current is called the Hall resistance, R_H.

Landau level (LL) The quantized kinetic energy of an electron in the presence of a magnetic field. The separation between Landau levels is called the cyclotron energy ($\hbar\omega_c$). The Landau levels of composite fermions are called CF-Landau levels.

Quantum fluid A fluid whose behavior is governed by quantum mechanical phases.

Quantum Hall effect (QHE) Occurrence of plateaus in the Hall resistance of a two-dimensional electron system on which it is quantized at $R_H = h/fe^2$, f being either an integer (the integral QHE) or a fraction (the fractional QHE). The plateau with $R_H = h/fe^2$ is centered at filling factor $\nu = f$.

Two-dimensional electron system (2DES) A system of electrons confined to two dimensions. Such a system is obtained typically at the interface of two semiconductors.

von Klitzing constant (*R*$_K$) $R_K \equiv h/e^2$.

I. QUANTUM HALL EFFECT

The study of magnetotransport in systems where electrons are confined to move in two dimensions has given us one of the most fascinating phenomena discovered in physics: the quantum Hall effect. Its theoretical investigation has helped uncover new structures and concepts, and there is a consensus that a sound understanding of the basic physics of this new quantum fluid has been achieved. This article makes an attempt to summarize in a simple, coherent, and least redundant manner the generally accepted knowledge at the present, to which a large number of workers have contributed. I apologize to those whose work could not be adequately represented due to length constraints or my ignorance; the books edited by R. E. Prange and S. M. Girvin (1990), S. Das Sarma and A. Pinczuk (1996), and O. Heinonen (1998) ought to be consulted for a comprehensive bibliography as well as a more detailed historical account.

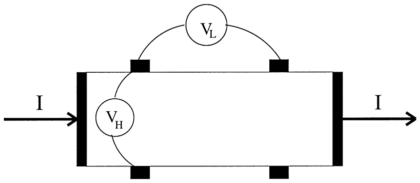


FIGURE 1 Schematics of magnetotransport measurements. I, V_L, and V_H are the current, longitudinal voltage, and the Hall voltage, respectively. The longitudinal and Hall resistances are defined as $R_L \equiv V_L/I$ and $R_H \equiv V_H/I$.

In 1879, E. H. Hall discovered that the passage of current in the presence of a magnetic field induces a voltage perpendicular to the direction of the current flow, an effect known as the Hall effect (see Fig. 1). A new resistance, known as the Hall resistance, is defined as

$$R_H = \frac{V_H}{I}. \tag{1}$$

The phenomenon can be understood in simple classical terms, based on the Lorentz force law of electrodynamics, which tells us that the Hall resistance is given by

$$R_H = \frac{B}{\rho e c}, \tag{2}$$

where B is the external field, and ρ is density of current carriers. The proportionality of the Hall resistance to B is used routinely to measure the density of the mobile charges.

The modern field of quantum Hall effect (QHE) began almost exactly 100 years later, when K. von Klitzing in 1980, and D. C. Tsui, H. L. Stormer, and A. C. Gossard found in 1982 that in two-dimensional electron systems, the Hall resistance exhibits plateaus on which its value is quantized (Fig. 2), determined only by universal constants of nature:

$$R_H = \frac{h}{fe^2}, \tag{3}$$

where f can be either an integer (the integral quantum Hall effect, or IQHE) or a fraction (the fractional quantum Hall effect, or FQHE). Concomitant with the quantization of R_H is an exponential suppression of the longitudinal resistance R_L with temperature, indicating a total lack of dissipation in the limit of zero temperature.

The absolute accuracy of the quantization of R_H has been established to 2.4 parts in 10^8 (for one standard deviation uncertainty), and the relative accuracy to 3.5 parts in 10^{10} at National Institute of Standards and Technology (Jeffery *et al.*, 1998) and the Swiss Federal Office of Metrology (Jeckelman *et al.*, 1995). The quantization is believed to be exact. The ratio h/e^2 has been adopted as the fundamental unit of resistance, called the von Klitzing constant (R_K), with its value given by $R_K = 25812.807572(95)$ Ohms. The combination h/e^2 also occurs in the definition of the fine structure constant $\alpha = e^2/\hbar c \approx 1/137$. Because the speed of light is known extremely precisely, the Hall effect measurements in dirty solid-state systems also provide one of the most accurate values for α.

The appearance of a universal quantization, independent of the sample type, geometry, and various materials parameters (like the band mass of the electron or the dielectric constant of the semiconductor), caught the community by surprise. One might expect such physics in, say, a simple atomic system, but it was entirely unexpected in a complex, macroscopic, and disordered solid-state system. Simple behaviors in complex systems have always enthralled physicists, and the discovery of QHE predictably stimulated intense activity in search of the fundamental principles responsible for it.

II. INTEGRAL QUANTUM HALL EFFECT

The IQHE, namely, the quantization of the Hall resistance at $R_H = h/ne^2$ where n is an integer, was discovered by von Klitzing in 1980. That the phenomenon has a quantum mechanical origin (hence the Q in QHE) was obvious by the appearance of the Planck's constant in the formula for the Hall resistance. Our understanding of the IQHE, discussed later, shows that it is a single particle effect, that is, it can be understood in an independent electron model. It is a dramatic consequence of the well-known quantization of the electron kinetic energy into Landau levels (LLs).

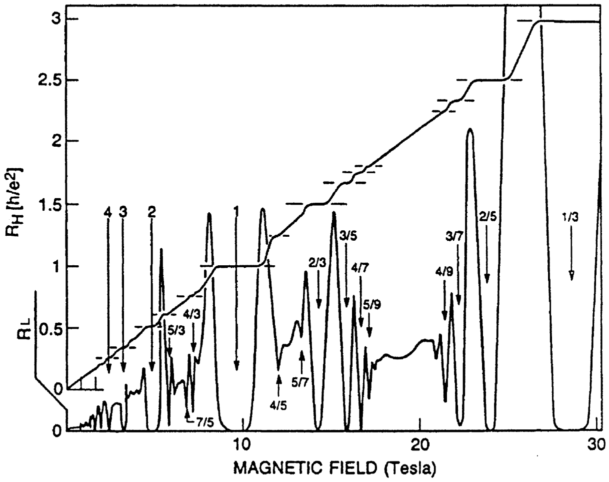


FIGURE 2 Overview of the quantum Hall effect. The Hall resistance R_H and the longitudinal resistance R_L are plotted as a function of the magnetic field B. (Adapted from Stormer and Tsui, 1996.)

A. Landau Levels

The Hamiltonian for a single electron moving in two dimensions in a perpendicular magnetic field is given by

$$H = \frac{1}{2m_b}\left(\mathbf{p} + \frac{e}{c}\mathbf{A}\right)^2, \tag{4}$$

where, $\mathbf{A}$ is given by $\nabla \times \mathbf{A} = \mathbf{B} = B\hat{\mathbf{z}}$ and m_b is the band mass of the electron. The solution of this problem shows that the electron kinetic energy is quantized into Landau levels, as shown in Fig. 3B,C. The eigenenergies are given by $E_n = \hbar\omega_c(n + 1/2)$, where $\omega_c = eB/m_bc$ is the cyclotron frequency, and $n = 0, 1, \ldots$ is the Landau level index. The degeneracy of each Landau level for a single spin is given by B/ϕ_0 per unit area, where $\phi_0 = hc/e$ is called the flux quantum.

Now consider many independent electrons, with density ρ per unit area. The ground state is obtained by filling up the lowest energy single particle orbitals, with the condition that no orbital is occupied by more than one electron, as required by the Pauli principle. The number of filled Landau bands,

$$\nu = \frac{\rho\phi_0}{B} \tag{5}$$

is called the filling factor (defined so that a Landau level filled with both spins has a filling factor of 2). The plateau with $R_H = h/ne^2$ is centered at $\nu = n$.

The many-particle ground state is infinitely degenerate in general, because all arrangements of electrons in the topmost partially filled Landau level produce the same energy. The exception is at an integral filling factor, $\nu = n$. The ground state here is unique (Fig. 3B), with a gap to excitations. This fact is responsible for the IQHE plateau with $R_H = h/ne^2$. The IQHE is a consequence of the quantization of the electron energy into Landau levels.

B. Disorder and QHE Plateaus

Even though the physics of the IQHE lies in the opening of a gap at integral filling factors, it turns out that a finite amount of disorder is also needed for the establishment of the plateaus. It is easy to see that no plateaus may result in the absence of disorder. Consider the motion of electrons in crossed electric and magnetic fields in a system without disorder. Taking advantage of the translational invariance of the problem, we can boost to a frame of reference moving with velocity $v = cE/B$ in which there is no electric field, and hence no current. This allows a calculation of the current in the laboratory frame of reference, which yields the classical value of the Hall resistance, Eq. (2), with no plateaus.

The Landau levels are broadened by disorder as depicted schematically in Fig. 4, with extended states at the

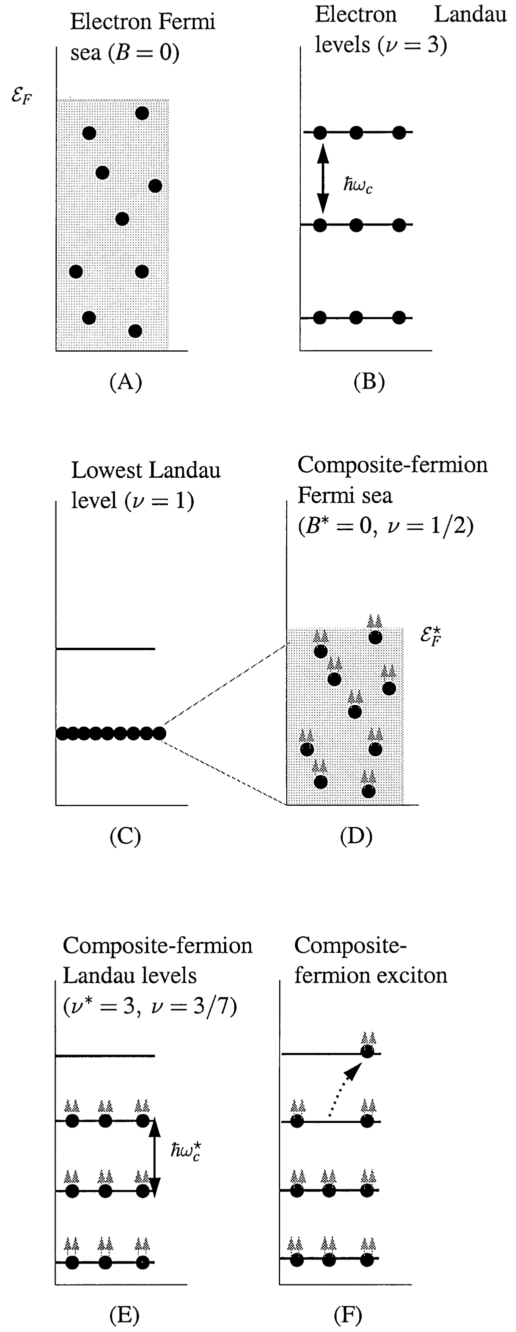

Electron Fermi sea ($B = 0$)
$\mathcal{E}_F$
Electron Landau levels ($\nu = 3$)
$\hbar\omega_c$
(A)
(B)
Lowest Landau level ($\nu = 1$)
Composite-fermion Fermi sea ($B^* = 0,\ \nu = 1/2$)
$\mathcal{E}_F^*$
(C)
(D)
Composite-fermion Landau levels ($\nu^* = 3,\ \nu = 3/7$)
Composite-fermion exciton
$\hbar\omega_c^*$
(E)
(F)

FIGURE 3 Evolution of two-dimensional electron system as the transverse magnetic field B is increased. For independent electrons, the Fermi sea (A) (filled to Fermi energy $\mathcal{E}_F$) at $B = 0$, splits into Landau levels (B). The lowest Landau level (C) is split by interactions into energy levels of composite fermions. The composite fermions are shown as electrons with attached vortices, with each vortex represented by an arrow. These fill a CF-Fermi sea (D) at $n = 1/2$ (filled to Fermi energy $\mathcal{E}_F^*$) and occupy CF-LLs (E) at other filling factors. A jump out of such a level (F) creates a CF exciton. At still higher fields, the scenario (D–F) repeats itself, but now with composite fermions carrying four or more flux quanta. The particle spin has been neglected for simplicity of illustration. (Adapted from Jain, 2000.)

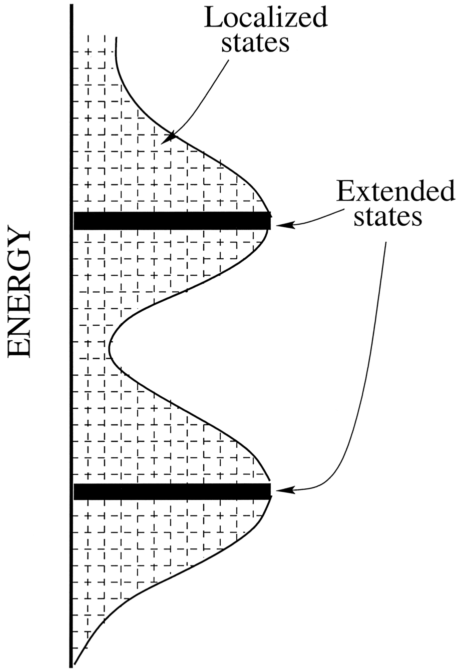


FIGURE 4 Schematics of the single particle density of states as a function of energy in the presence of disorder. The Landau levels are broadened, with extended states at the centers and localized states elsewhere.

centers and localized states elsewhere. A somewhat oversimplified description of the physics of plateaus is as follows. At integral filling $\nu = n$, the classical formula Eq. (2) for the Hall resistance can be recast to give $R_H = h/ne^2$. Now imagine changing the filling factor away from $\nu = n$ by adding some electrons or holes to the system. So long as the additional particles go into orbitals that are localized, they do not contribute to transport; as a result, the Hall resistance retains its value $R_H = h/ne^2$.

A more general and precise explanation for the Hall plateau follows from an argument due to R. B. Laughlin (1981), which showed that the Hall resistance is quantized at $R_H = h/ne^2$ whenever the Fermi level lies in the localized states, with n counting the number of extended bands below the Fermi level. Consider a Hall bar with periodic boundary conditions, which has the topology of the Corbino disk, shown in Fig. 5. Imagine that ideal, disorder-free regions have been attached at the inner and the outer boundaries of the Corbino sample. An azimuthal current I flows when a voltage V_H is applied across the sample. The current I is related to the variation in the energy of the system as a function of a test flux ϕ through the center by $I = c(dU/d\phi)$, which is approximated by $I = c(\Delta U/\phi_0)$, where ΔU is the change in energy during the process of changing the test flux adiabatically by $\phi_0 = hc/e$. Because a flux quantum through the center can be gauged away, the single particle energy levels at $\phi = 0$ and $\phi = \phi_0$ are identical. The energy U changes because the occupations of the single particle energy levels are different at $\phi = \phi_0$ than at $\phi = 0$. The determination of ΔU requires monitoring the evolution of single particle orbitals as the test flux ϕ is varied from 0 to ϕ_0. Each of the extended states, defined as states that go around the sample thereby encircling the test flux, moves to the next one during this process. The localized states, which do not enclose the test flux, remain unaffected. Let us now consider the situation when all of the extended states of the lowest n LLs are occupied, i.e., the Fermi level lies in localized states. As the extended states evolve under the variation of ϕ they carry their electrons with them, because there is nowhere else for the electrons to go. At the end of the adiabatic process, each extended state has moved into the next one, carrying its electron with it, with the net effect that precisely n electrons have been transported from the inner to the outer edge of the Corbino disk. We therefore have $\Delta U = neV_H$ which gives us the quantized Hall resistance $R_H = V/I = h/ne^2$.

A closely related approach for understanding the IQHE is based on the Landauer-Büttiker theory of resistance (Büttiker, 1988). This approach is especially useful for dealing with deviations from exact quantization. Consider a sample with translational invariance in the x direction, so p_x, the momentum in the x direction, is a good quantum number. The potential in the y direction is a combination of the confinement potential at the edges, and the potential

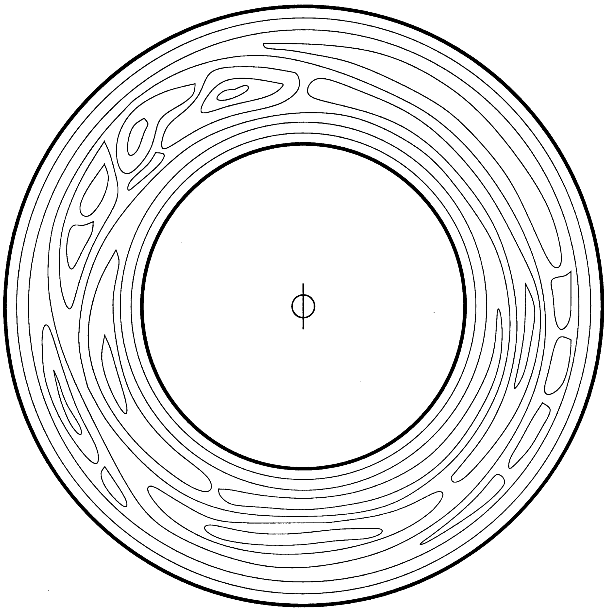


FIGURE 5 Schematic depiction of the single particle states in the Corbino geometry, with a test flux f piercing through the hole.

due to the applied Hall voltage. Let us now consider a full Landau level, that is, a state in which all single particle energy levels with momenta $p_- < p_x < p_+$ are occupied, where p_- and p_+ are the momenta at the edges of the sample. The current in the x direction is obtained by adding the individual currents:

$$I = e\int \frac{dp_x}{2\pi\hbar} v_x. \tag{6}$$

With $v_x = (1/m_b)\langle p_x + (e/c)A_x\rangle = \langle \partial H/\partial p_x\rangle = \partial E/\partial p_x$, it follows that $I = (e/h)(\mu_+ - \mu_-) = (e^2/h)V_H$, where μ_+ and μ_- are the chemical potentials at the edges. The current is independent of the LL index, so the total current for n filled Landau levels is $I = (ne^2/h)V_H$. Now consider a sample with disorder, but connected to the current source via ideal, disorder-free leads. So long as the electrons at the chemical potential, which are at one edge of the sample, are not back-scattered, the current is not degraded and the Hall resistance remains quantized at $R_H = h/ne^2$, independent of disorder. The back-scattering is strongly suppressed because the states carrying currents in opposite directions are localized on opposite edges, and therefore are exponentially weakly coupled.

III. FRACTIONAL QUANTUM HALL EFFECT

A. Phenomenology

In 1982, Tsui, Stormer, and Gossard observed a plateau quantized at $R_H = h/\frac{1}{3}e^2$. In the subsequent years, as the result of a tremendous improvement in the quality of samples, a host of new fractions were observed. Today, the number of observed fractions below unity $(f < 1)$ stands at approximately 50 and increasing. The plateau at $R_H = h/fe^2$ is seen in the vicinity of filling factor $\nu \approx f$. The longitudinal resistance R_L exhibits activated behavior, as in IQHE, indicating the existence of a gap in the excitation spectrum. The observed fractions appear in sequences of the form

$$f = \frac{n}{2pn \pm 1}. \tag{7}$$

Some of the fractions observed to date are:

$$f = \frac{n}{2n+1} = \frac{1}{3}, \frac{2}{5}, \frac{3}{7}, \cdots \frac{10}{21}$$

$$f = \frac{n}{2n-1} = \frac{2}{3}, \frac{3}{5}, \frac{4}{7}, \cdots \frac{10}{19}$$

$$f = \frac{n}{4n+1} = \frac{1}{5}, \frac{2}{9}, \cdots \frac{6}{25}$$

$$f = \frac{n}{4n-1} = \frac{2}{7}, \frac{3}{11}, \cdots \frac{6}{23}.$$

The first several members of each sequence are well established, in that quantized plateaus have been observed. The last few are seen only through resistance minima, but there is little doubt that the corresponding Hall plateaus will develop upon further improvement in sample quality. FQHE at f also implies FQHE at $1 - f$ due to particle-hole symmetry in the lowest Landau level. In addition to these fractions, FQHE has also been observed with $f = 5/2$, the only exception to the "odd-denominator rule" in a single layer system.

B. Model Hamiltonian

Because only the integral QHE is possible for independent electrons, interactions are clearly responsible for producing gaps at fractional filling factors. One therefore must look for the solutions of the more complete problem of interacting electrons, defined by the Schrödinger equation $H\Psi = E\Psi$ with

$$H = \sum_j \frac{1}{2m_b}\left[\frac{\hbar}{i}\nabla_j + \frac{e}{c}\mathbf{A}(\mathbf{r}_j)\right]^2 + \frac{e^2}{\epsilon}\sum_{j<k}\frac{1}{|\mathbf{r}_j - \mathbf{r}_k|} + \sum_j U(\mathbf{r}_j) + g\mu\mathbf{B}\cdot\mathbf{S}. \tag{8}$$

The first term on the right-hand side is the kinetic energy, the second term is the Coulomb interaction energy, the third term is a one-body potential incorporating the effects of the uniform positive background and disorder, and the last term is the Zeeman energy. The parameter ϵ is the dielectric constant of the background material.

Insofar as the conceptual foundation of the FQHE is concerned, it is convenient to neglect disorder, and consider the limit of large B, so both the cyclotron and the Zeeman energies are large compared to the interaction energy. The electrons are then fully polarized and confined to the lowest LL, making the kinetic and Zeeman energies irrelevant constants. We thus end up with the idealized model of fully polarized electrons in the lowest LL with the Hamiltonian (suppressing the interaction with the background)

$$H_{LLL} = \frac{e^2}{\epsilon}\sum_{j<k}\frac{1}{|\mathbf{r}_j - \mathbf{r}_k|}. \tag{9}$$

This is the simplest and the cleanest model containing the essential physics of the FQHE. The strongly coupled, nonperturbative nature of the problem can already be seen by noting that there is no small parameter in the problem: H_{LLL} contains only one energy scale, set by the Coulomb interaction. All states are degenerate in the absence of interaction, and the FQHE results as soon as the interaction is turned on, no matter how small its strength (or, how large ϵ). Standard perturbative treatments are not useful here.

The FQHE is a true many body phenomenon, in which strongly interacting electrons behave in a correlated manner to produce rich and nontrivial, yet amazingly simple behavior. The solution of this Schrödinger equation should clarify the physics responsible such behavior. By analogy to the IQHE, the plateau at $R_H = h/fe^2$ with fractional f originates due to the opening of a gap at $\nu = f$. (For such a gapped state, when the filling factor is moved away from $\nu = f$, some "defects" are created, but they are pinned by disorder, thus giving rise to a plateau at $R_H = h/fe^2$.) The goal of theory is therefore to explain the origin of gaps in a partially filled Landau level, specifically why they appear only at certain sequences of odd-denominator fractions. A satisfactory explanation of the odd-denominator rule must necessarily elucidate why there is no FQHE at even-denominator fractions (with the exception of $f = 5/2$). Even though the problem looks intractable at first, the trail of experimental clues has led to a simple, yet extremely accurate solution to the problem that is also in good agreement with experimental observations.

C. Laughlin's Theory

The first observed fraction was $f = 1/3$. Soon thereafter, in 1983, Laughlin noted that the single particle wave function in lowest Landau level has the form $z^s \exp[-|z|^2/4l^2]$, where $z = x - iy$ denotes the position of the electron as a complex number, and $l = \sqrt{\hbar c/eB}$ is the magnetic length. The wave function of a system containing many electrons must therefore have the form $F_S[\{z_j\}] \exp[-\sum_k |z_k|^2/4l^2]$ where F_S is an antisymmetric polynomial of z_j. Choosing a Jastrow form for the polynomial, he wrote the following wave function

$$\Psi_{1/m} = \prod_{j<k} (z_j - z_k)^m \exp\left[-\frac{1}{4l^2} \sum_i |z_i|^2 \right], \qquad (10)$$

which provides an excellent representation of the ground state at $\nu = 1/m$, as confirmed by comparison with exact results known for small systems. This wave function served as a paradigm for the subsequent theoretical developments. It is easy to see that it has good correlations built in it in the presence of repulsive interactions. In a typical wave function satisfying the Pauli principle, the probability of two electrons approaching one another vanishes as r^2, r being the distance between them. In $\Psi_{1/m}$, it vanishes much faster, as r^{2m}, which shows that electrons avoid each other efficiently.

The exponent in the Jastrow factor, m, must be an odd integer, due to the fundamental requirement of the antisymmetry of the wave function. For $m = 1$, $\Psi_{1/m}$ gives the wave function of the fully occupied lowest LL ($\nu = 1$), and for $m = 5$ it gives the wave function at $\nu = 1/5$, which is relevant to FQHE at $f = 1/5$ observed later on. However, the exponent is not allowed to take noninteger or even-integer values, and the observations of numerous fractions other than $1/m$ subsequent to Laughlin's theory pointed to the existence of a more general structure.

D. Composite Fermions

A more general theory was put forth by the author in 1989. The fundamental building block of this theory is called the composite fermion, which is the bound state of an electron and an even number of quantum mechanical vortices. According to this theory, strongly interacting electrons in the lowest LL capture vortices to turn into weakly interacting composite fermions. The ensuing investigations revealed that the FQHE was only one manifestation of composite fermions, which describe a superstructure encompassing other phenomena as well. Experimenters observed their Fermi sea, their Shubnikov-de Haas oscillations, and their semiclassical cyclotron orbits; they also measured the particles' charge, spin, statistics, mass, and magnetic moment (Stormer and Tsui, 1996).

The composite fermion (CF) theory proposes the following wave functions at any arbitrary filling factor ν:

$$\Psi_\nu = \Phi_{\nu^*} \prod_{j<k=1}^{N} (z_j - z_k)^{2p}, \qquad (11)$$

where Φ_{ν^*} are the known wave functions of noninteracting electrons at an effective filling factor ν^*, related to ν as

$$\nu = \frac{\nu^*}{2p\nu^* \pm 1}. \qquad (12)$$

This equation gives wave functions for ground as well as low-energy excited states of interacting electrons at arbitrary ν, derived from the corresponding states of noninteracting electrons at ν^*. Ψ_ν describe a strongly correlated state of electrons, because the probability of electrons approaching one another in Ψ_ν vanishes rapidly as r^{4p+2}. In the limit of very strong B, the functions Ψ are to be projected into the lowest electronic LL.

The wave functions Ψ in Eq. (11), which are accurate representations of the actual eigenstates (see below), lend themselves to a simple interpretation. They contain a Jastrow factor $\prod_{j<k}(z_j - z_k)^{2p}$ which attaches $2p$ vortices to each electron. The bound state of an electron and $2p$ vortices behaves as a particle, called the composite fermion. The electronic wave function Ψ can therefore also be interpreted as wave functions of composite fermions.

The capture of vortices has a profound consequence for the dynamics of the particles. As composite fermions move about, the vortices carried by them generate phases which partly cancel the Aharonov-Bohm phases due to the external magnetic field, and the composite fermions in effect experience a much weaker magnetic field. Because a closed path around a vortex produces, by definition, a phase of 2π, a vortex is effectively equivalent to a flux quantum, which also produces the same Aharonov Bohm phase for a closed path around it. (For this reason, the composite fermion is often envisioned as an electron bound to $2p$ flux quanta.) Therefore, each vortex cancels one flux quantum of the external magnetic field, giving the effective magnetic field:

$$B^* = B - 2p\rho\phi_0. \qquad (13)$$

In effect, each electron absorbs $2p$ flux quanta of the external field to become a composite fermion (Fig. 6). (It must be understood here that an external magnetometer will measure B and not B^*. There is no real "Meissner effect" in the FQHE. However, B^* is the real magnetic field for composite fermions; this is the field that would be obtained if the composite fermions themselves are used to measure the field.) Treating composite fermions as independent, their filling factor is $\nu^* = \rho\phi_0/|B^*|$, and Eq. (13) can be transcribed into the relation in Eq. (12). The $-$ sign in Eq. (12) corresponds to the situation when B^* points opposite to B.

Laughlin's theory of inverse-odd-integer states falls naturally within the CF theory. The wave function $\Psi_{1/(2p+1)} = \Phi_1 \prod_{j<k}(z_j - z_k)^{2p}$ is identical to Laughlin's wave function, which can be seen by noting that the wave function of the fully occupied lowest LL is given by

$$\Phi_1 = \prod_{j<k}(z_j - z_k)\exp\left[-\frac{1}{4l^2}\sum_i |z_i|^2\right]. \qquad (14)$$

$\Psi_{1/(2p+1)}$ is interpreted as one filled Landau level of composite fermions.

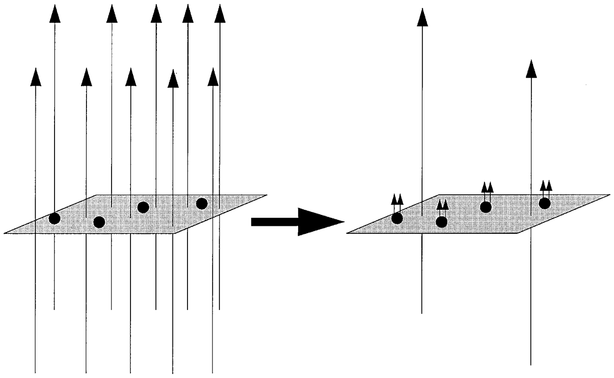


FIGURE 6 Capturing two flux quanta converts each electron into a composite fermion moving in a reduced effective magnetic field. (Adapted from Jain, 2000.)

The unusual character of composite fermion ought to be emphasized. It is a collective particle, with the definition of one composite fermion involving all particles in the system. A composite fermion may live only inside the CF liquid. It is an inherently quantum mechanical object, a "quantum particle," because it is the product of the union of an electron and quantum mechanical phases (vortices). The fluids of composite fermions are quantum fluids not only because composite fermions themselves are quantum particles, but also because they involve a quantization of the composite fermion orbits into CF Landau levels. The composite fermion also has a topological character due to its integrally quantized vorticity ($=2p$). It is indeed surprising that the composite fermions behave as ordinary particles to a large degree.

In 1991, A. Lopez and E. Fradkin implemented the physics of composite fermions in a Chern-Simons field theoretical framework. It has been further developed by a number of groups over the years. The Hamiltonian formulation of R. Shankar and G. Murthy (1997) obtains many essential features at the Hartree-Fock level.

To summarize: The strongly interacting electrons in the lowest Landau level transforms into weakly interacting composite fermions in a reduced magnetic field B^*. The lowest electronic Landau level thus splits into energy levels of composite fermions (Fig. 3).

E. FQHE

Plotting the FQHE data as a function of B^* brings out its striking similarity to the IQHE data plotted as a function of B, as seen in Fig. 7. The difference between the two is no more significant than that between the IQHE traces in different samples demonstrating that the strongly correlated liquid of interacting electrons at B behaves as a weakly interacting system of fermions at B^*. A similar correspondence is obtained for negative values of B^*, and also at lower electron filling factors, where composite fermions have four or more vortices bound to them.

There is also a correspondence between the Hall plateaus at B and B^*, but the Hall resistances are not the same at ν^* and ν. For example, at $\nu = 1/2$, where $B^* = 0$, the Hall resistance is $R_H = h/(\frac{1}{2}e^2)$, but at $B = 0$ we have $R_H = 0$. The difference is explained by noting that the composite fermions respond to a combination of the external Hall voltage and the Hall voltage induced by the vortex current tied to the charge current in the CF state, but the latter is not measured by the external voltmeter. (Just like the effective magnetic field, the induced Hall electric field is also fully internal to composite fermions.)

The FQHE of electrons is understood as the IQHE for composite fermions. The integral fillings $\nu^* = n$ of composite fermions correspond to electron filling factors

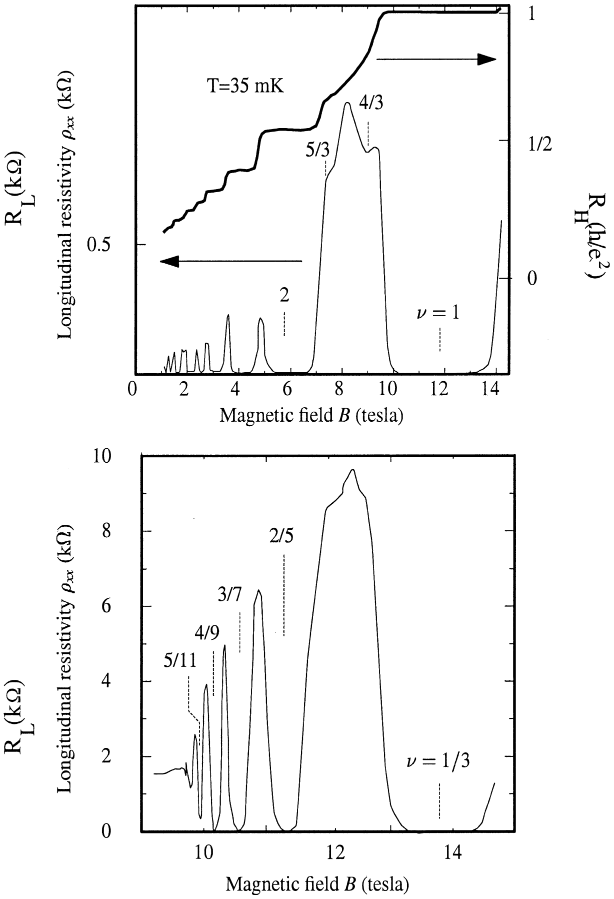


FIGURE 7 The top panel shows the IQHE of electrons. The bottom panel shows the FQHE of electrons plotted as a function of B^*, starting from $B^* = 0$ (n = 1/2). A close correspondence between the prominent features is manifest. (Adapted from Clark, 1986, and Du, 1993.)

$\nu = n/(2pn \pm 1)$, which are precisely the observed fractions. The composite fermion is to the FQHE what the electron is to the IQHE. Just as the IQHE is an observation of the electron Landau levels, the FQHE is an observation of the composite-fermion Landau levels. The phenomena of the IQHE and the FQHE, which were at first thought to be distinct, thus turn out to be intimately related: They are both integral quantum Hall effects, but for different particles. This unification of the FQHE and the IQHE is not surprising in view of the empirical similarity between the two (Fig. 2).

F. Computer Experiments

For a finite number of particles, the Hilbert space in lowest Landau level is finite, allowing a complete and exact solution of the problem through a brute force numerical diagonalization of the Coulomb Hamiltonian. Even though the systems are finite, they are sufficiently large (10–15 particles) to provide the opportunity for rigorous, detailed, and nontrivial tests of the theory. The computer experiments are also cleaner than the real experiments, in that the idealized limits of no disorder and large magnetic field can be explicitly implemented. For these reasons, computer experiments have played an important role in the theory of the FQHE. Figure 8 shows a number of exact spectra for interacting electrons in the lowest Landau level at the special filling factors of $\nu = 1/3$, 2/5, and 3/7. The spherical geometry (Haldane, 1983) is used in these calculations, which considers electrons moving on the surface of a sphere with a radial magnetic field of appropriate strength through the surface of the sphere; the total orbital angular momentum L is used to label the eigenstates.

The energy spectrum predicted by the CF theory, obtained without any adjustable parameters, is shown by dots

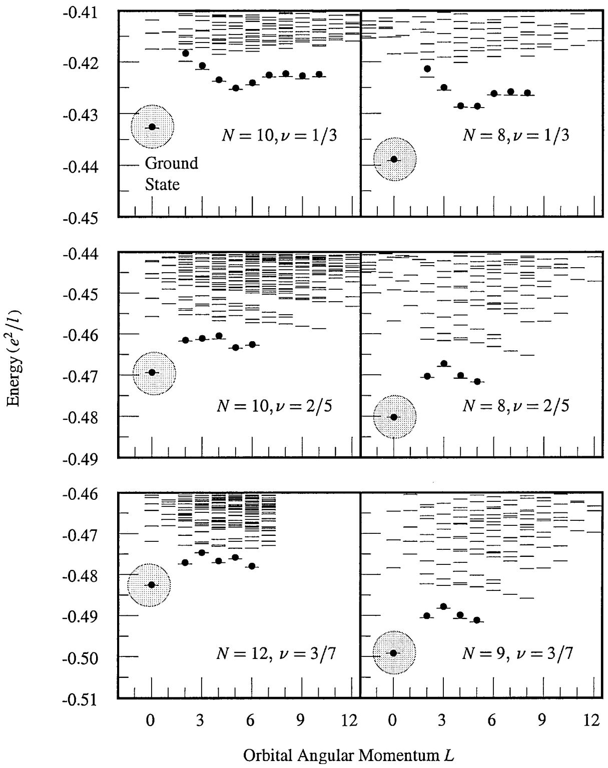


FIGURE 8 Energy spectrum for systems with $N = 8$–12 iteracting electron at n = 1/3, 2/5 and 3/7 moving on the surface of a sphere in the presence of a radial magnetic field. The dashes show the exact eigenenergies and the dots show the CF predictions obtained without any adjustable parameters. L is the total orbital angular momentum. The ground state (encircled) is described as an integer number of filled CF-LLs, and the branch of low-lying excited states, decorated with dots, represents the CF-exciton in various possible configurations. The energy is quoted in units of e^2/l, where l is the magnetic length. (Adapted from Jain and Kamilla, 1998.)

in Fig. 8. The CF energies agree with the exact eigenenergies to within 0.05–0.1%, and the overlaps between the exact eigenstates and the corresponding CF wave functions are close to perfect (typically >99%) for all systems studied. Numerous such studies have confirmed the description of the FQHE ground state as the state containing n-filled CF-LLs, and the lowest energy branch of excitations as the CF-exciton (a particle-hole pair of composite fermions).

G. Excitations and CF Mass

The quantitative understanding of real experiments is less accurate than that of computer experiments, because the experimental numbers are unavoidably affected by the nonzero transverse thickness of the wave function (the dynamics is still strictly two dimensional because only the lowest transverse subband is occupied), Landau level mixing, and disorder, all conveniently set to zero in the computer experiments. As a result, the experimental results themselves vary from sample to sample, depending on various parameters such as the form of the confinement potential, electron density, band mass, or mobility. An incorporation of these sample-specific effects into theory requires approximations which, even in the best cases, introduce uncertainties on the order of 20–30% in the theoretical numbers.

The dispersion of the neutral CF exciton (Fig. 3F) contains, in general, several minima, called rotons. The roton at $\nu = 1/(2p+1)$ was first obtained by S. M. Girvin, A. H. MacDonald, and P. M. Platzman in 1985 in a single-mode approximation theory. The roton energies have been determined experimentally for several FQHE states in both Raman and ballistic phonon scattering experiments, and are in good agreement (10–20%) with the theoretical calculations of V. W. Scarola and co-workers (2000). The neutral excitation in the long wavelength limit was observed by A. Pinczuk and co-workers in 1993 in Raman experiments (Fig. 9).

The longitudinal resistance displays activated behavior, $R_L \propto \exp[-\Delta_a/2k_BT]$, over a range of temperature. The activation energy Δ_a is identified with the energy of a far separated particle-hole pair of composite fermions. The agreement between theory and experiment is somewhat worse in this case (~factor of 2), presumably because the energy of the charged excitation is more strongly affected by disorder, neglected in theory. Δ_a is interpreted as the effective cyclotron energy for composite fermions, $\hbar eB^*/m^*c$, which defines a composite fermions mass, m^* (Du, 1993). A reasonably consistent interpretation of the activation energies for different FQHE states is obtained in terms of a single mass parameter. For typical parameters, the mass of the composite fermion is comparable to the electron mass in vacuum and approximately an order of magnitude larger than the band mass of the electron in GaAs, but unrelated to either; it is completely determined by the interaction between electrons.

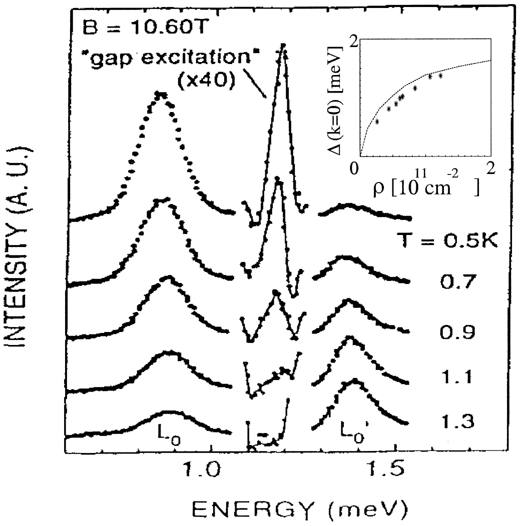


FIGURE 9 The long-wavelength collective mode (labeled the "gap excitation") at n = 1/3 in Raman scattering. The inset shows a comparison between the experimental data (stars) (Kang *et al*., 2001) and the theoretical estimation of two-roton bound state energy (dashed line). Theoretical estimates (Park and Jain, 2000) are obtained by considering Landau level mixing as well as finite thickness effects. (Adapted from Pinczuk, A. *et al*. 1993.)

H. Fractional Charge

Laughlin showed in 1983 that the charge of the excitation in a FQHE state is fractionally quantized. At $\nu = n/(2pn \pm 1)$, the value of the charge is $|e^*| = e/(2pn \pm 1)$, as seen most simply from the following argument. Take the state at $\nu = n/(2pn \pm 1)$ and adiabatically insert a flux through the origin from zero to ϕ_0. For a nondegenerate state, this creates a vortex (or an antivortex depending on the direction of the flux), the charge of which can be shown to be of magnitude νe, as follows essentially from the definition of the filling factor. The vortex is in general a collection of an integral number (r_1) of "elementary" excitations, i.e., $r_1|e^*| = [n/(2pn \pm 1)]e$. On the other hand, because an added electron must decay into elementary excitations, we must also have $r_2 e^* = e$, r_2 being another integer. The two conditions are compatible only with fractional values for e^*, the largest solution being $|e^*| = e/(2pn \pm 1)$. This derivation clarifies that the fractional charge is an inescapable, model-independent consequence of the existence of a nondegenerate ground state in a partially filled LL, and its valued is fixed by the filling factor. In the CF theory, the fractional charge of the

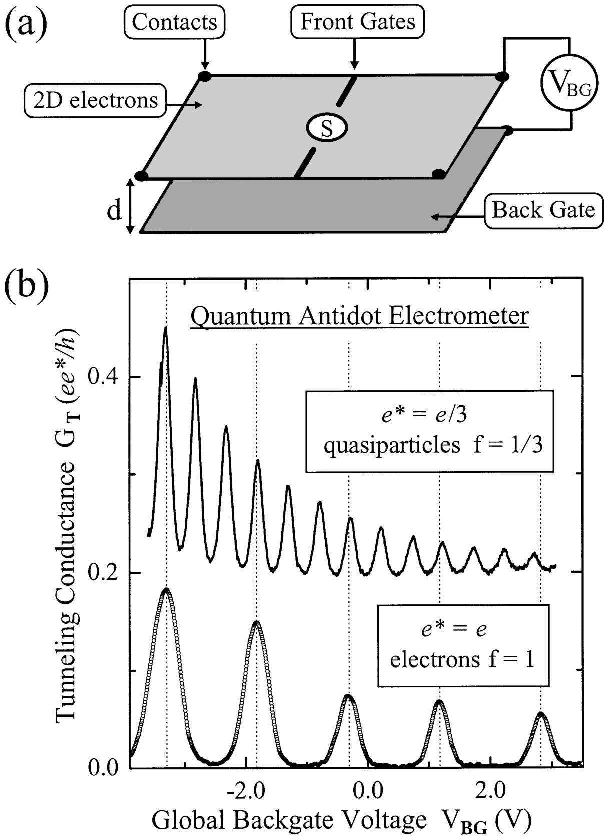


FIGURE 10 Charge is added on to a quantum antidot S by variation of magnetic field, and measured by capacitive coupling to a back gate parallel to the 2DES. Each conductance peak in resonant tunneling through the antidot indicates the addition of one unit of charge. A factor of three difference between the periods in the IQHE (lower curve) and the FQHE (n = 1/3, upper curve) demonstrates $e^* = e/3$ for the latter. (Adapted from Goldman, 2000.)

composite fermion excitation is most simply obtained by noting that the charge associated with it is given by the charge of the electron, $-e$, plus the charge deficiency due to $2p$ vortices, $2pe\nu$.

The fractional charge was measured in 1995 by V. J. Goldman and B. Su in resonant tunneling experiments (Fig. 10), and in 1997 by L. Saminadayar and co-workers and R. de Picciotto and co-workers in shot noise experiments.

I. Spin Physics

At sufficiently large B, when the Zeeman splitting $E_Z \to \infty$, the low-energy states are maximally polarized, and the spin of the electron is frozen. However, E_Z is quite small for typical experimental parameters (Halperin, 1983). Due to a small band mass (0.07 of the free electron mass) and a small g factor (-0.44 as opposed to 2 in vacuum), the ratio $E_Z/\hbar\omega_c$ is only $\sim 1/70$ in GaAs. E_Z is also much smaller than the typical interaction energy. This raises the possibility that the ground state may not always be maximally polarized.

A naive application of the Hund's rule of atomic physics would lead to a maximally polarized ground state even for $E_Z = 0$. However, the application of the Hund's rule to composite fermions rather than electrons suggests a completely different picture. The electron state at $\nu = n/(2pn \pm 1)$ is mapped into $\nu^* = n$ of composite fermions as before, but now we have $n = n_\uparrow + n_\downarrow$, where $n_\uparrow$ ($n_\downarrow$) is the number of spin-up (spin-down) CF Landau bands occupied. At zero Zeeman energy, the ground state is a spin singlet when n is an even integer (with $n_\uparrow = n_\downarrow = n/2$), and partially spin polarized for odd n (with $n_\uparrow = (n+1)/2$ and $n_\downarrow = (n-1)/2$). Upon increasing E_Z, the ground state spin changes in discontinuous jumps when the CF-LLs of up and down spins cross. These qualitative considerations were confirmed in the extensive experiments of R. R. Du and co-workers in 1995, in which they varied E_Z by application of an additional magnetic field parallel to the 2D layer. The results are well described by an effective mass model for composite fermions (Fig. 11). These

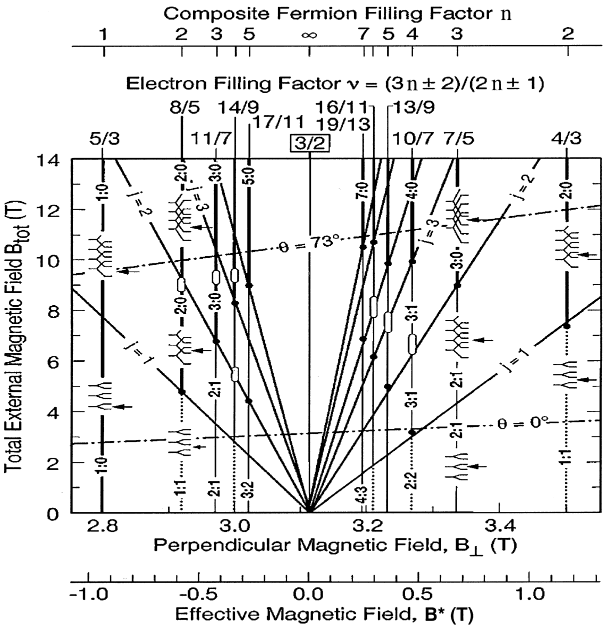


FIGURE 11 The positions (dots) of transitions between states with different spin polarizations as a function of the Zeeman energy (y-axis) at various fillings (x-axis) around n = 3/2 (The filling factors n = $(3n \pm 2)/(2n \pm 1)$ are related to n = $n/(2n \pm 1)$ by particle-hole symmetry, which relates n to 2 − n for spinful electrons). The CF-LL occupation is shown pictorially in each region, labeled by $n_\uparrow : n_\downarrow$. The solid lines emanating from the origin are from a model of independent composite fermions with the CF mass and g factor treated as adjustable parameters. (Adapted from R. R. Du *et al.* (1995).)

observations verify that the spin of the composite fermion is 1/2.

J. Even Denominator Fractions

At $\nu = 1/2$, the simplest even-denominator fraction, the effective field vanishes for composite fermions. V. Kalmeyer and S. C. Zhang (1992), and B. I. Halperin, P. A. Lee, and N. Read (1993) formulated the metallic state here in terms of a Fermi sea of composite fermions. At precisely $\nu = 1/2$, composite fermions move in straight lines. Slightly away from $\nu = 1/2$, where B^* is very small, they are expected to execute semiclassical cyclotron orbits. The radius of the cyclotron orbit of a composite fermion at the Fermi surface is given by $R^* = \hbar k_F^*/eB^*$, with $k_F^* = \sqrt{4\pi\rho}$, as appropriate for a fully polarized Fermi sea. R^* involves only known parameters, and can be orders of magnitude of larger than any electronic length scale in the problem. The cyclotron radius of the charge carriers was measured in 1993–1994 through magnetic focusing (Fig. 12), geometrical antidot resonances (Fig. 13), and surface-acoustic-wave attenuation measurements (Willett *et al.*, 1993), and was found to be in agreement with R^* of the CF theory. Besides demonstrating the existence of composite fermions outside of the FQHE (because no FQHE is seen in the vicinity of $\nu = 1/2$), these experiments also explicitly confirmed the fermionic nature of composite fermions through the observation of their Fermi sea. The Shubnikov-de Haas oscillations, thermoelectric power, and spin polarization measurements are also consistent with the composite-fermion Fermi sea description. The absence of FQHE at $\nu = 1/2$ is explained because the Fermi sea has no gap to excitations.

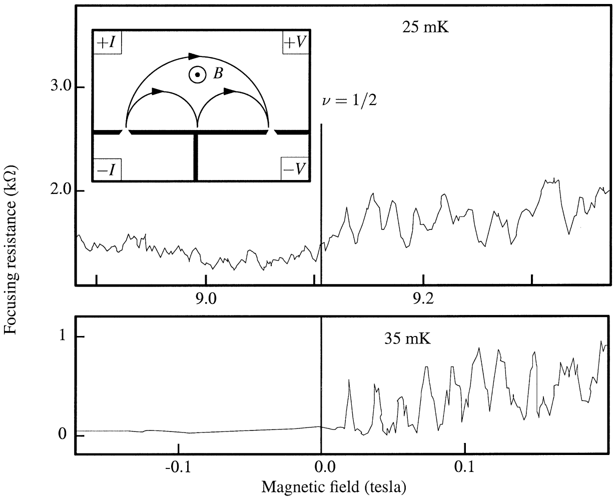


FIGURE 12 Direct determination of the effective magnetic field by magnetic focusing of composite fermions by injecting them into one constriction and collecting into another, possibly after an integer number of bounces (inset). The lower panel shows the focusing peaks for *electrons* at $B \approx 0$, while the upper depicts the focusing peaks for composite fermions with $B^* \approx 0$. The focusing peaks (superimposed over mesoscopic resistance fluctuations due to disorder) align after scaling B^* by a factor of $\sqrt{2}$ to account for the fact that the electron Fermi sea is spin unpolarized whereas the composite-fermion Fermi sea is spin polarized. (Adapted from Goldman *et al.*, 1994.)

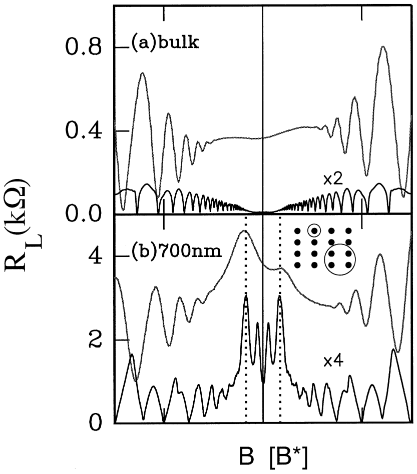


FIGURE 13 The upper panel shows the usual magnetoresistances. The lower panel shows geometric resonances in an antidot superlattice (inset) for electrons (lower curve) and composite fermions (upper curve). The two pairs of peaks for electrons near the origin correspond to two smallest cyclotron orbits of electrons commensurate with the antidot lattice, shown in the inset. The two broad peaks for composite fermions corresponds to the resonance due to the smallest CF cyclotron orbit. The x-axis is the real magnetic field for electrons and the effective magnetic field for composite fermions. (For comparison, B^* has been scaled by $\sqrt{2}$ as in Fig. 12.) (Adapted from Kang, W., *et al*. (1993).)

K. Exactness of Hall Quantization

The principle governing the exactness of the Hall quantization can be traced to the topological feature that the number of vortices bound to each electron must be exactly quantized to be an integer, as required by the fundamental principle of single-valuedness of the quantum-mechanical wave function. This results in an exactly quantized quasiparticle charge, which in turn guarantees the exactness of the Hall quantization, following Laughlin's 1981 argument that relates the quasiparticle charge to the value of the Hall resistance. The odd-denominator rule is a consequence of the antisymmetry of Ψ, which requires $2p$ to be an *even* integer. The Hall quantization is thus a macroscopic manifestation of microscopic postulates of quantum mechanics. (It should be emphasized that the accuracy of the wave functions Ψ does not by itself imply the exactness of the Hall quantization, but it gives us confidence that the *physics* of binding of $2p$ vortices to each electron is exact.)

IV. MORE PHENOMENA

Many other phenomena have been investigated in the context of the QHE. In all cases, there has been a healthy interaction between theory and experiment to produce rapid progress.

A. 2D Localization in Magnetic Field

As discussed earlier, localization of states is crucial for the establishment of plateaus. This has motivated an intense investigation into the nature of single particle states in the QHE regime in the presence of disorder, and much progress has been made (Das Sarma and Pinczuk, 1996). Numerical solution of the Schrödinger equation makes a strong case that truly extended states occur at only one energy (E_c) in each Landau band, with the localization length diverging as $\xi \sim |E - E_c|^{-\gamma}$ with $\gamma \approx 2.3$. The width of the transition region between two plateaus is experimentally found to vanish as $T^{-0.42}$ with temperature; the value of the exponent is consistent with the theoretical value $1/\gamma z \approx 0.43$, provided the dynamical exponent is taken to be $z = 1$, as expected for quantum phase transitions in Coulomb systems.

B. Wigner Crystal

When the interaction energy dominates over the kinetic energy, it is believed that electrons form a lattice called the Wigner crystal (WC). Because the kinetic energy is effectively suppressed in the lowest LL, one might *a priori* have expected a WC here rather than the FQHE. A variational calculation, reproduced in Fig. 14, shows that the CF liquid has a significantly lower energy than the WC for a range of ν, but the WC wins at sufficiently small ν. There is good experimental evidence for an insulating behavior at small $\nu (< 1/5)$, interpreted as a pinned Wigner crystal (Jiang, 1990); observation of nonlinear I–V (Goldman, 1990) supports this view.

C. Skyrmion

The neglect of interactions at integral fillings is valid in the $B \to \infty$ limit, but interactions may affect the nature of excitations significantly under typical experimental conditions. In 1993, S. L. Sondhi, A. Karlhede, S. A. Kivelson, and E. H. Rezayi showed theoretically that the excitation at $\nu = 1$ is typically not a simple spin reversed electron but has a nontrivial spin texture, described in a (nonlinear sigma-model) field theory as a skyrmion. It was observed in 1995 by S. E. Barrett and coworkers. The size of the skyrmion shrinks rapidly with increasing Zeeman energy, but skyrmions with as many as 30 reversed spins have been observed in systems where the g factor was suppressed by application of hydrostatic pressure.

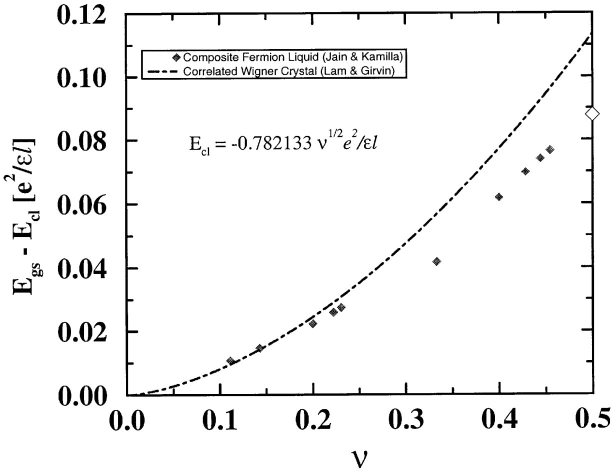


FIGURE 14 The variational energies per particle of the CF liquid and the Wigner crystal as a function of the filling factor. $E_{cl} = -0.782133\sqrt{\mathrm{n}}e^2/el$ is the energy of a classical two-dimensional Wigner crystal with triangular symmetry. The open diamond on the right vertical axis is the estimate of the CF Fermi sea energy at $\mathrm{n} = 1/2$, obtained by an extrapolation. The energy of the CF liquid is shown only at the special $n/(2pn+1)$ filling factors; the full curve will have downward cusps at these points. (Source: Jain and Kamilla, 1998; Lam and Girvin, 1984.)

D. Edge States

Even when there is a gap to excitations in the bulk, gapless excitations exist at the edges of the sample. It is believed that the physics of the edges is effectively one dimensional, described by the well-developed theory of Tomanaga-Luttinger liquids. In 1990 X. G. Wen argued that the quantized value of the Hall resistance fixes the parameters of the Luttinger model, leading to definite predictions for various power laws, say, for the resistance associated with tunneling into and out of the edges of a FQHE system through a weak link. Significant progress toward experimental tests of some aspects of theory has been made, notably by A. M. Chang and co-workers.

E. Pairing of Composite Fermions

At filling factor $\nu = 5/2 = 2 + 1/2$, the lowest Landau level is fully occupied and the filling factor is 1/2 in the second Landau level. Treating the electrons in the lowest LL as inert, one would naively expect a Fermi sea of composite fermions in the second Landau level. However, a FQHE was observed here already in 1987 by Willett and co-workers. There is growing support to the view that the FQHE originates here due to a pairing of composite fermions, which opens up a gap. Variational and exact diagonalization studies indicate that the paired CF state is described by a particle-hole symmetrized version of a Pfaffian wave function written in 1991 by G. Moore and N. Read.

F. Stripes

The abundance of FQHE in the lowest LL is in contrast to its near absence in higher LLs. Very few fractions are seen in the second LL ($2 < \nu < 4$) and none whatever in third and higher LLs, indicating that some other physics takes over in higher Landau levels. In a Hartree-Fock calculation, A. A. Koulakov, M. M. Fogler, and B. I. Shklovskii showed in 1996 that close to $\nu = n + 1/2$ in high LLs the system prefers to phase separate into alternating stripes of $\nu = n$ and $\nu = n + 1$. Observation of a strongly anisotropic R_L at $\nu \approx n + 1/2$ for $n > 4$ (M. Lilly and co-workers, 1999; R. R. Du and co-workers, 1999) supports this picture.

G. Multilayer Systems

Two far-separated layers are simply two single-layer systems. However, when the layers are sufficiently close, new structure may arise. An interesting example is the FQHE at $\nu = \frac{1}{2} = \frac{1}{4} + \frac{1}{4}$, observed by Y. W. Suen *et al.* and J. P. Eisenstein *et al.* in 1992. It is well described by a multicomponent wave function of Halperin (1983).

H. Topological Considerations

In 1985, to explain the precision of the Hall quantization, Niu, Thouless, Wu, and Kohmoto showed that the value of the quantized Hall resistance in the integral QHE is related to a topological invariant called the Chern number. Thouless *et al.* (1982) also considered QHE in periodic geometries, for which the Landau level splits into Hofstadter bands, and showed that the Hall resistance of each band is integrally quantized, with the integer characterizing the quantization depending sensitively on the value of flux per plaquette.

ACKNOWLEDGMENTS

The author was supported in part by the National Science Foundation. He thanks V. W. Scarola, K. Park, and M. Marder for help with figures.

SEE ALSO THE FOLLOWING ARTICLES

Electrodynamics, Quantum • Electron Spin Resonance • Quantum Mechanics

BIBLIOGRAPHY

Barrett, S. E., *et al.* (1995). “Optically Pumped NMR Evidence for Finite-Size Skyrmions in GaAs Quantum Wells near Landau Level Filling $\nu = 1$,” *Phys. Rev. Lett.* **74,** 5112–5115.

Büttiker, M. (1998). “Absence of backscattering in the quantum Hall effect in multiprobe conductors,” *Phys. Rev.* **B38,** 9375–9389.

Chang, A. M., *et al.* (2001). “Plateau Behavior in the Chiral Luttinger Liquid Exponent,” *Phys. Rev. Lett.* **86,** 143–146.

Clark, R. G., *et al.* (1986). “Odd and even fractionally quantized states in GaAs-GaAlAs heterojunctions,” *Surf. Sci.* **170,** 141–147.

Das Sarma, S., and Pinczuk, A. (eds.) (1996). “Perspectives in Quantum Hall Effects,” Wiley, New York.

De Picciotto, R., *et al.* (1997). “Direct observation of a fractional charge,” *Nature* **389,** 162–164.

Du, R. R., *et al.* (1993). “Experimental evidence for new particles in the fractional quantum Hall effect,” *Phys. Rev. Lett.* **70,** 2944–2947.

Du, R. R., *et al.* (1995). “Fractional quantum Hall effect around $\nu = 3/2$: Composite fermions with a spin,” *Phys. Rev. Lett.* **75,** 3926–3929.

Du, R. R., *et al.* (1999). “Strongly anisotropic transport in higher two-dimensional Landau levels,” *Solid State Commun.* **109,** 389–394.

Eisenstein, J. P., *et al.* (1992). “New fractional quantum Hall state in double-layer two-dimensional electron systems,” *Phys. Rev. Lett.* **68,** 1383–1386.

Girvin, S. M., MacDonald, A. H., and Platzman, P. M. (1985). “Collective Excitation Gap in the Fractional Quantum Hall Effect,” *Phys. Rev. Lett.* **54,** 581–583.

Goldman, V. J., Santos, M., Shayegan, M., and Cunningham, J. E. (1990). “Evidence for two-dimensional quantum Wigner crystal,” *Phys. Rev. Lett.* **65,** 2189–2192.

Goldman, V. J., Su, B., and Jain, J. K. (1994). “Detection of composite fermions by magnetic focusing,” *Phys. Rev. Lett.* **72,** 2065–2068.

Goldman, V. J. (2000). “The Quantum Antidot Electrometer: Direct Observation of Fractional Charge,” *J. Korean Phys. Soc.*, in press.

Haldane, F. D. M. (1983). “Fractional Quantization of the Hall Effect: A Hierarchy of Incompressible Quantum Fluid States,” *Phys. Rev. Lett.* **51,** 605–608.

Halperin, B. I. (1983). “Theory of the quantized Hall conductance,” *Helv. Phys. Acta* **56,** 75–86.

Halperin, B. I., Lee, P. A., and Read, N. (1983). “Theory of the half-filled Landau level,” *Phys. Rev.* **B47,** 7312–7343.

Heinonen, O. (ed.) (1998). “Composite Fermions,” World Scientific, New York.

Jain, J. K. (1989). “Composite Fermion Approach for the Fractional Quantum Hall Effect,” *Phys. Rev. Lett.* **63,** 199–202.

Jain, J. K., and Kamilla, R. K. (1998). Chapter 1 in Heinonen, O. (ed.) “Composite Fermions,” World Scientific, New York.

Jain, J. K. (2000). “The Composite Fermion: A Quantum Particle and Its Quantum Fluids,” *Physics Today* **53**(4), 39–45.

Jeckelmann, B., Inglis, A. D., and Jeanneret, B. (1995). “Material, Device, and Step Independence of the Quantized Hall Resistance,” *IEEE Transactions on Instrumentation and Measurement,* **44,** 269–272.

Jeffery, A., *et al.* (1998). “Determination of the von Klitzing constant,” *Metrologia* **35,** 83–96.

Jiang, H. W., *et al.* (1990). “Quantum liquid versus electron solid around $\nu = 1/5$ Landau-level filling,” *Phys. Rev. Lett.* **65,** 633–636.

Kalmeyer, V., and Zhang, S. C. (1992). “Metallic phase of the quantum Hall system at even-denominator filling fractions,” *Phys. Rev.* **B46,** 9889–9892.

Kang, M., *et al.* (2001). “Observation of Multiple Magnetorotons in the Fractional Quantum Hall Effect,” *Phys. Rev. Lett.* **86,** 2637–2640.

Kang, W., Stormer, H. L., Pfeiffer, L. N., and West, K. W. (1993). “How real are composite fermions?” *Phys. Rev. Lett.* **71,** 3850–3853.

Kohmoto, M. (1985). "Topological Invariant and the Quantization of the Hall Conductance," *Ann. Phys.* **160,** 343–353.

Koulakov, A. A., Fogler, M. M., and Shklovskii, B. I. (1996). "Charge Density Wave in Two-Dimensional Electron Liquid in Weak Magnetic Field," *Phys. Rev. Lett.* **76,** 499–502.

Lam, P. K., and Girvin, S. M. (1984). "Liquid-solid transition and the fractional quantum-Hall effect," *Phys. Rev.* **B30,** 473–475.

Laughlin, R. B. (1981). "Quantized Hall conductivity in two dimensions," *Phys. Rev.* **B23,** 5632–5633.

Laughlin, R. B. (1983). "Anomalous Quantum Hall Effect: An Incompressible Quantum Fluid with Fractionally Charged Excitations," *Phys. Rev. Lett.* **50,** 1395–1398.

Lilly, M. P., *et al.* (1999). "Evidence for an Anisotropic State of Two-Dimensional Electrons in High Landau Levels," *Phys. Rev. Lett.* **82,** 394–397.

Lopez, A., and Fradkin, E. (1991). "Fractional quantum Hall effect and Chern-Simons gauge theories," *Phys. Rev.* **B44,** 5246–5262.

Moore, G., and Read, N. (1991). "Nonabelions in the fractional quantum Hall effect," *Nucl. Phys.* **B360,** 360–396.

Niu, Q., Thouless, D. J., and Wu, Y. S. (1985). "Quantized Hall conductance as a topological invariant," *Phys. Rev.* **B31,** 3372–3377.

Park, K., and Jain, J. K. (2000). "Two-Roton Bound State in the Fractional Quantum Hall Effect," *Phys. Rev. Lett.* **84,** 5576–5579.

Pinczuk, A., Dennis, B. S., Pfeiffer, L. N., and West, K. (1993). "Observation of collective excitations in the fractional quantum Hall effect," *Phys. Rev. Lett.* **70,** 3983–3986.

Prange, R. E., and Girvin, S. M. (eds.) (1990). "The Quantum Hall Effect," Springer-Verlag, New York.

Saminadayar, L., Glattli, D. C., Jin, Y., and Etienne, B. (1997). "Observation of the $e/3$ fractionally charged Laughlin quasiparticle," *Phys. Rev. Lett.* **79,** 2526–2529.

Scarola, V. W., Park, K., and Jain, J. K. (2000). "Rotons of composite fermions: Comparison between theory and experiment," *Phys. Rev.* **B61,** 13064–13072.

Shankar, R., and Murthy, G. (1997). "Towards a Field Theory of Fractional Quantum Hall States," *Phys. Rev. Lett.* **79,** 4437–4440.

Sondhi, S. L., Karlhede, A., Kivelson, S. A., and Rezayi, E. H. (1993). "Skyrmions and the crossover from the integer to fractional quantum Hall effect at small Zeeman energies," *Phys. Rev.* **B47,** 16419–16426.

Stormer, H. L., and Tsui, D. C. (1996). Chapter 10 in Das Sarma, S., Pinczuk, A. (eds.) "Perspectives in Quantum Hall Effects," Wiley, New York.

Suen, Y. W., *et al.* (1992). "Observation of a $\nu = 1/2$ fractional quantum Hall state in a double-layer electron system," *Phys. Rev. Lett.* **68,** 1379–1382.

Thouless, D. J., Kohmoto, M., Nightingale, M. P., and den Nijs, M. (1982). "Quantized Hall Conductance in a Two-Dimensional Periodic Potential," *Phys. Rev. Lett.* **49,** 405–408.

Tsui, D. C., Stormer, H. L., and Gossard, A. C. (1982). "Two-Dimensional Magnetotransport in the Extreme Quantum Limit," *Phys. Rev. Lett.* **48,** 1559–1562.

Von Klitzing, K., Dorda, G., and Pepper, M. (1980). "New Method for High-Accuracy Determination of the Fine-Structure Constant Based on Quantized Hall Resistance," *Phys. Rev. Lett.* **45,** 494–497.

Wen, X. G. (1990). "Chiral Luttinger Liquid and the Edge Excitations in the Fractional Quantum Hall States," *Phys. Rev.* **B41,** 12828–12844.

Willett, R. L., Ruel, R. R., West, K. W., and Pfeiffer, L. N. (1993). "Experimental demonstration of a Fermi surface at one-half filling of the lowest Landau level," *Phys. Rev. Lett.* **71,** 3846–3849.

Willett, R. L., *et al.* (1987). "Observation of an even-denominator quantum number in the fractional quantum Hall effect," *Phys. Rev. Lett.* **59,** 1776–1779.

Quantum Mechanics

Albert Thomas Fromhold, Jr.
Auburn University

GLOSSARY

Eigenfunction Wavefunction for a specific stationary state of a physical system.

Eigenvalue Specific value of a physical quantity corresponding to a specific eigenfunction.

Eigenvalue equation Particular mathematical relation in which a mathematical form, known as an operator, acts on an eigenfunction to produce the corresponding eigenvalue multiplied by the eigenfunction.

Excited states States of a system having energies above the ground state.

Exclusion principle Statement that no two identical particles of a particular statistical type can simultaneously occupy the same quantum state.

Ground state Lowest energy state of a system.

Hermitian operator A mathematical operator having real eigenvalues, which is a necessary condition for it to be capable of representing a physical observable.

Lifetime Mean time before an excited state spontaneously decays to another state, such as the ground state.

Particlelike Localized and acting in an individual manner as an entity.

Probability density Relative probability of a particle being at a specified position in space; alternately, the relative probability of some other physical quantity, such as momentum.

Quantization Discrete value or set of values for a physical quantity, such as energy or angular momentum.

Stationary state Specific state of a physical system characterized by the fixed value of some physical quantity, such as energy.

Time-dependent Schrödinger equation Equation for determining the time-dependence and space-dependence of the eigenfunctions for a system.

Time-independent Schrödinger equation Eigenvalue equation used to obtain energy eigenvalues and energy eigenfunctions for a system.

Uncertainty relation Mathematical form relating the maximum precision of measurement of some physical quantity, such as position or energy, to the precision of some related quantity, such as momentum or time.

Wavefunction A mathematical form, usually complex, used to deduce the probability density.

Wavelike Nonlocalized and periodic, with the capability of interacting constructively or destructively.

Wave–particle duality Coexistence of wavelike and particlelike aspects in a physical entity, such as an electron.

QUANTUM MECHANICS is a theory that is capable of predicting the behavior of atomic and subatomic systems. In fact, it is the only theory to date that is adequate for the microscopic domain in nature. Quantum mechanics not only correctly predicts the results of physical observations in the microscopic world where classical physics is often quite unsuccessful but also leads to valid predictions in the macroscopic world experienced by human senses.

I. CLASSICAL MODEL OF THE ATOM

A. Structure of the Atom

Ernest Rutherford's (1871–1937) interpretation of his extensive scattering experiments in 1911 gave overpowering evidence that atoms consist of a dense, positively charged nucleus surrounded by a cloud of electrons. The electron had been discovered a few years earlier in 1887 by Joseph John Thomson (1856–1940), who attributed a definite charge-to-mass ratio to the particle. Before the discovery of the wave-like properties of the electron in 1927 by George Paget Thomson (1892–1975), the son of J. J. Thomson, it was expected that the mechanical properties of atoms could readily be explained by applying classical mechanics in a straightforward way to this model. In fact, the successful model of planetary motion around the more massive sun, under the action of gravitational forces between the planets and the sun, and the perturbations to such motion due to the gravitational forces between planets, provide the closely related classical analog model for describing the motion of electrons about a single, massive nucleus: the electrical forces between charged particles replace the gravitational forces in the solar system.

The forces between electrons are repulsive instead of attractive, and these repulsive forces are of the same order of magnitude as the attractive forces between an electron and the nucleus. Therefore these forces between electrons are not merely perturbations, as is the case for the gravitational forces between planets orbiting the sun.

However, it was expected that the hydrogen atom, containing only a single electron and thus free of the repulsive Coulomb force between electrons in two-electron and many-electron atoms, should be easily amenable to treatment by classical mechanics. The model is a simple picture in which the single electron orbits the far more massive proton nucleus under the action of a centripetal force provided by the attractive electrical force between electron and nucleus.

B. Classical-Mechanical Treatment of the Single-Electron Ion

Let us consider the more general case of a single electron atom or ion, where the charge of the nucleus is Ze, with the quantity e representing the magnitude of the electron charge. For the hydrogen atom, $Z = 1$. The electrical force between a nucleus of charge Ze separated by a distance r from an electron of charge $-e$ has magnitude

$$F_{\mathrm{E}} = \frac{KZe^2}{r^2}, \tag{1}$$

where K is a constant that for SI units has the value

$$K = \frac{1}{4\pi\varepsilon_0}, \tag{2}$$

where ε_0 is the electric permittivity of free space having the value 8.854×10^{-12} farad/meter (F/m). Let us make the assumption here that the nucleus is stationary and the electron travels around the nucleus. Although this view is absolutely correct in classical mechanics only if the nucleus is infinitely massive, it can be shown to be approximately correct whenever the nucleus is much more massive than the electron, and this is satisfied for all one-electron ions. Thus, we consider the radius of the electron orbit to be equal to the distance separating the electron and the nucleus; corrections for deviations caused by the finite nuclear mass can easily be incorporated into the treatment at a later stage. The electrical force provides the centripetal force mv^2/r, which causes the electron to travel in the hypothesized circular orbit, m being the electron mass and v being the electron speed, so that

$$\frac{mv^2}{r} = \frac{KZe^2}{r^2}. \tag{3}$$

This at once gives the following nonrelativistic expression for the kinetic energy $\mathscr{E}_{\mathrm{K}}$ of the electron:

$$\mathscr{E}_{\mathrm{K}} = \frac{1}{2}mv^2 = \frac{KZe^2}{2r}. \tag{4}$$

The potential energy $\mathscr{E}_{\mathrm{P}}$ associated with the Coulomb force is given by

$$\mathscr{E}_{\mathrm{P}} = \frac{-KZe^2}{r}, \tag{5}$$

so that

$$\mathscr{E}_{\mathrm{K}} = -\frac{1}{2}\mathscr{E}_{\mathrm{P}}. \tag{6}$$

The sum of potential and kinetic energies gives the total energy $\mathscr{E}_{\mathrm{T}}$:

$$\mathscr{E}_{\mathrm{T}} = \mathscr{E}_{\mathrm{P}} + \mathscr{E}_{\mathrm{K}} = -2\mathscr{E}_{\mathrm{K}} + \mathscr{E}_{\mathrm{K}} = -\mathscr{E}_{\mathrm{K}}. \tag{7}$$

The kinetic energy is intrinsically positive, as can be noted from Eq. (4), so that the total energy is negative. This means that the electron is bound to the nucleus. As the electron separation from the nucleus increases, the potential energy algebraically increases toward zero. If the electron cannot separate to an arbitrarily large distance from the nucleus, we say that it is bound to the nucleus.

The centripetal force relation given by Eq. (3) relates the electron speed v to the radius r of the classical orbit, so that

$$r = \frac{KZe^2}{mv^2} \tag{8}$$

or, equivalently

$$v = \left(\frac{KZe^2}{mr}\right)^{1/2}. \tag{9}$$

The rotation angular frequency ω is given by

$$\omega = \frac{v}{r} = \left(\frac{KZe^2}{mr^3}\right)^{1/2}. \tag{10}$$

This development not only allows the kinetic energy to be expressed in terms of the radius, as given by Eq. (4), it also allows the angular momentum $L = pr = mvr$ for an electron in a circular orbit to be expressed in terms of the radius,

$$L = pr = mvr = mr\left(\frac{KZe^2}{mr}\right)^{1/2} = (KZe^2mr)^{1/2}. \tag{11}$$

Equivalently, this gives the radius in terms of the angular momentum

$$r = \frac{L^2}{KZe^2m}, \tag{12}$$

which can then be substituted into Eq. (9) to give the speed in terms of the angular momentum

$$v = \left[\left(\frac{KZe^2}{m}\right)\left(\frac{KZe^2m}{L^2}\right)\right]^{1/2} = \frac{KZe^2}{L}. \tag{13}$$

The kinetic energy in terms of the angular momentum is then given by

$$\mathscr{E}_{\mathrm{K}} = \frac{1}{2}mv^2 = \frac{mK^2Z^2e^4}{2L^2} \tag{14}$$

Equations (1)–(14) are based entirely on classical mechanics.

C. Electromagnetic Fields Produced by an Orbiting Electron

Let us explore the classical viewpoint a bit further by means of the planetary model of the one-electron atom, where a single electronic charge $-e$ is considered in motion about an equal-magnitude charge of opposite sign due to the proton. This constitutes a rotating electric dipole, although admittedly the center of rotation is located very near one end of the dipole. This rotating dipole produces a time-varying electric field at distant points in space. The electric field so produced is cyclic, being the same as the rotation frequency of the dipole. An oscillating electric field is thus produced at the observation point, the frequency of the field ν being equal to the frequency of rotation of the dipole. According to classical electrodynamics, the energy associated with this oscillating electric field can propagate outward in space or be absorbed at the field point (e.g., by accelerating the conduction electrons in a metal).

D. Electromagnetic Radiation Predictions of Classical Physics

The energy radiated away or absorbed from a rotating dipole source, as previously described, must come at the expense of the potential energy of the configuration of the two charges if it is not supplied by some source connected with the rotation of the dipole. For a hydrogen atom, there is no such source. The consequent decrease in total energy of the electron in the one-electron atom would mean that the total energy becomes more negative, which corresponds to an increase in the kinetic energy of the electron in accordance with Eq. (7). This corresponds to a decrease in the radius of the electron orbit and to an increase in the rotation frequency according to Eqs. (4) and (10). On the basis of this classical model of radiation, the frequency of the radiated electromagnetic wave would then increase. The changes which occur would thus be continuous; that is, since there could be arbitrary values for the radius of each electron orbit, the electron would be capable of having any one of a continuous range of values of kinetic energy. This speed could be changed continuously by adding or extracting arbitrarily small quantities of energy. As the speed and frequency of rotation change, corresponding continuous changes occur in the radius of the orbit. According to classical electrodynamics, an accelerated charge radiates energy, so the potential energy of the electron would steadily decrease. This ultimately leads to a catastrophe: there would be theoretically no limit to the process, even as the total energy of the electron–proton system approached negative infinity, with the consequence that an infinite amount of energy would be radiated away.

E. Logical Failure of Classical Mechanics

Clearly, the results deduced on the basis of the classical model are unreasonable. In addition, the predictions do not correspond in any way to the experimental observations of optical spectra, which are not continuous, but instead contain many sharp spectral lines. Hence, this classical model cannot be used to describe the mechanical properties of an atom.

The true state of affairs is that all atoms have a lowest energy state, labeled the *ground state* by Niels Henrik Bohr (1885–1962), for which no further energy emission is possible. Furthermore, even while in some given higher-energy configuration, the atom does not emit energy continuously; excited states emit energy only sporadically, each such event being accompanied by a sudden jump to a lower-energy configuration. Bohr called the various discrete energy states of an atom stationary states, since such states are stable until the time when a transition to a lower-energy state actually takes place.

The logical deduction from the previous discussion only can be that the classical approach, which is so successful for describing the motion of the planets, fails completely for the hydrogen atom. So much for arguing by analogy! Moreover, the failure of classical mechanics for the hydrogen atom is manifested not in our lack of ability to see and follow the electron, because it might be too tiny to be seen with the eye or even with a microscope, but instead in experimental measurements yielding data such as the light spectrum emitted by excited gases of atoms. The observed spectra can in no way be explained without additional assumptions quite foreign to classical mechanics.

Even considered statistically, an attractive, inverse square force, as typified by Eq. (1) for the Coulomb electric force, leads to a negative potential energy that varies inversely with the separation distance between charges (or of masses, in the case of gravitational forces). The potential energy is conservative, meaning that a decrease in separation distance yields energy that can be extracted by an external agency or converted into heat. If the separation distance can be imagined to decrease to zero, then the potential energy approaches negative infinity. This implies that an infinite amount of energy simultaneously would either be extracted or be converted into heat. Conceptually, all machines in the world could be run for a day, a month, a year, or even as long as one wishes, by allowing an electron to come closer and closer to a proton in a controlled manner. This means that in the original creation of the opposite charges, an infinite amount of energy was expended. Alternately, we could imagine an explosion dwarfing even that of the most powerful fission or fusion weapon available today coming about by merely allowing or triggering the collapse of a configuration of two point charges of opposite sign. Such incomprehensible conclusions resulting from pushing the classical model to its logical endpoint are sufficient in themselves to force us into the realization that nature actually must behave under constraints additional to those contained within the formalism of classical mechanics.

To bring in evidence bearing on this matter, electron–positron pairs can be produced from gamma rays, and it is known from these experiments on pair production that an infinite amount of energy is not required to separate the oppositely charged particles so produced. In fact, the γ-ray energy required is only of the order of the rest mass energies of the electron and the positron. Likewise, pair annihilation does not lead to the emission of an infinite quantity of γ-ray energy, so again we must conclude that the negative Coulomb energy must be restricted by nature to have a finite magnitude. This end could be accomplished by restricting the separation distance between opposite charges to some minimum, nonzero value corresponding, perhaps, to the electron–proton separation distance in the so-called ground-state configuration of the hydrogen atom.

F. New Foundations of Mechanics

It is important to note that classical mechanics does provide a generally adequate theoretical description of motion for all objects that can be seen in an ordinary way traveling at ordinary speeds. By "ordinary," we mean observable to the human eye and traveling at speeds low enough to be able to follow the position of the object with the eye. In fact, the theory is found to apply even in the domain of far smaller particles, such as can be seen only with the aid of an optical microscope, as long as the speed is well below that of light propagation. However, the extremely versatile framework provided by classical mechanics yields inaccurate results in two domains, which can be distinctly different: (i) particles traveling at speeds approaching the speed of light and (ii) particles having very tiny masses.

First of all, objects traveling at speeds v of the order of $0.1c$ or larger manifest marked departures from Isaac Newton's (1642–1727) predictions based on a fixed mass. This is due to the relativistic dependence of mass on velocity. Second, very small mass particles, such as the smaller atoms and nuclei (e.g., the hydrogen and helium atoms, proton, neutron, and α-particle) and especially the even less massive electrons, all exhibit diffraction properties typical of wavelike phenomena.

In both of these domains—namely, the domain of very fast particles and the domain of very tiny particles—entirely new forms of mechanics were erected that had essentially different premises than those inherent in classical mechanics. This was necessary because classical mechanics simply fails to describe and correctly predict the motions and trajectories of objects under such conditions. The domain listed in (i), very high-speed motion, requires the use of Albert Einstein's (1879–1955) theory of special relativity (1905). The domain listed in (ii), very small-mass particles, requires the use of the theory of quantum mechanics, the subject of this article.

Special relativity and quantum mechanics constitute a pair of theories that enable one to make accurate calculations in the separate domains listed. The *relativistic mechanics* provided by Einstein's theory of special relativity accurately describes and predicts motions of particles and bodies moving with speeds approaching the speed of light; quantum mechanics describes and predicts experimental observations for very tiny particles, atoms, and the properties of ensembles of atoms. Einstein's theory of special relativity alone, however, is unable to account for the properties of atoms and the discreteness of the emitted radiation that constitute the optically observed atomic spectra.

For very small-mass particles traveling at relativistic speeds, more complicated theories (relativistic quantum mechanics) have been formulated that have had their successes, one of the most prominent being Paul Adrien Maurice Dirac's (1902–1984) theory predicting the existence of the positron. Ideally, one would like to have a very general framework that not only predicted the motion of particles correctly for each of these domains but also gave the correct results under more ordinary conditions in which classical mechanics already does a completely satisfactory job. It must be admitted, however, that the basis for a perfectly general theory of mechanics, if it is within the province of human beings to conceive, still remains to be developed. Even Einstein's general theory of relativity does not contain the power to interpret the microscopic world but, instead, gives answers to cosmological questions relating to the macroscopic world. Simply put, no completely general theory exists today.

II. SPECIAL RELATIVITY

A. Essential Relations for Quantum Mechanics

Because special relativity is covered adequately in another part of this encyclopedia, only the relationships essential for the present development are presented here. It is vital to point out that the profound prediction by Einstein of the variation of the measured mass of an object with its speed has immediate implications for motion which are quite contrary to those predicted by Newton's second law, applied in the restricted sense of a fixed, unvarying mass for any given object. Contained within the theory of special relativity is not only the experimentally observed mass variation with speed but also the concept of the equivalence of mass and energy. The latter concept, which is directly verified in electron–positron pair production, has of course far-reaching consequences with regard to present-day energy production.

Defining the velocity $\mathbf{v}$ as a vector with magnitude equal to the speed and pointing in the direction of motion, the momentum $\mathbf{p}$ is then simply

$$\mathbf{p} = m\mathbf{v}. \tag{15}$$

Newton considered the mass m to have a fixed value m_0. The acceleration $\mathbf{a}$ is defined as the time rate of change of the velocity

$$\mathbf{a} = \frac{d\mathbf{v}}{dt}. \tag{16}$$

Newton's second law $\mathbf{F} = m_0\mathbf{a}$ thus can be written in terms of the time rate of change of the momentum

$$\mathbf{F} = \frac{d\mathbf{p}}{dt}. \tag{17}$$

This form is valid even in special relativity, as are Newton's first and third laws. Time t, however, is not absolute in special relativity as it is within the framework of Newton's formalism. Also, we know from special relativity that mass m depends upon speed in accordance with

$$m = \gamma m_0, \tag{18}$$

where m_0 is referred to as the *rest mass* and the parameter γ is determined by v/c, the ratio of the speed of the particle to the speed of light, in accordance with

$$\gamma = \left[1 - \left(\frac{v}{c}\right)^2\right]^{-1}. \tag{19}$$

Energy is related to mass m in special relativity:

$$\mathcal{E} = mc^2. \tag{20}$$

This relation holds true for any speed v. When $v = 0$, this reduces to the rest mass energy

$$\mathcal{E}_0 = m_0c^2. \tag{21}$$

The *energy versus momentum relation* for a free particle in special relativity is

$$\mathcal{E}^2 = \mathcal{E}_0^2 + p^2c^2, \tag{22}$$

which differs from the energy–momentum relation in Newtonian mechanics.

Energy–momentum relations are needed to develop the de Broglie relation, which underlies the Schrödinger equation of quantum mechanics.

B. Philosophical Implications

There is unquestionably a great philosophical difference between Einstein's special relativity and modern-day quantum mechanics. It is unlikely that Einstein could have foreseen in 1905, the year of publication of his works on special relativity and the photoelectric effect, the great philosophical differences that eventually would split the quantum way of thinking from the relativity way of thinking. Relativistic mechanics, like classical mechanics, leads to a world view of an absolutely predictable future, at least in principle, based on the specified state of the universe at any given time. Quantum mechanics, on the contrary, contains an intrinsic margin of uncertainty regarding the future evolution of the system, even if the state of the universe is known as completely as possible at any given time. For example, descriptions of the motion of an electron by Newtonian and Einsteinian mechanics are both quite precise, although not necessarily in agreement with each other, in predicting a specific position and velocity at a time t whenever the exact position and velocity at any other time t' are specified as well as all forces that act on the particle in the time interval between t' and t. Quantum mechanics, on the other hand, gives only relative probabilities for a broad range of possibilities.

One must be flexible enough to accept experimental facts and not reject, for example, relativistic mechanics simply because one dislikes the concept that the masses of objects vary with speed. In the same way, experimental demonstration of the wave properties of matter requires that theoretical formulations be broadened to include these properties. This leads in a natural way to the concept of indeterminism.

The lack of determinism inherent in quantum mechanics was the "Achilles heel" that eventually led Einstein to the conclusion that quantum mechanics provides, at best, an incomplete description of the universe. His powerful intuition, which should not be lightly discounted, led him to the belief that quantum mechanics more than likely would be superseded eventually by a more complete theory, completely deterministic in form, which would allow exact predictions of the complete future evolution of a system, given only the initial conditions at some prior point in time. This viewpoint, which means that nothing is left to chance, was hotly disputed by prominent contemporaries of Einstein. It would seem to belie even the possibility of free will in humans. Max Karl Ernst Planck (1858–1947) and Dirac, on the contrary, believed that the uncertainty inherent in quantum mechanics was in fact an intrinsic reality in nature, never to be overcome by any better theory purported to be more comprehensive or general. For these philosophically minded physical theorists, the important uncertainty principle is elevated from its role as an integral part of quantum mechanics theory to an even more lofty philosophical principle.

Philosophical speculation, of course, is not physics. One should ask different questions, such as how to best describe and predict experimental observations in a logical manner from models developed from a combination of observation and intuition. To go beyond this end by inquiring *why* nature behaves in such a fashion can prove interesting and stimulating but does not, in itself, constitute the furthering of the subject proper of physics.

The best approach is to develop formulations that represent a synthesis of all experimentally verified properties of matter, including variation of mass with velocity and wavelike interference of particles. The roots of the most fundamental relation (the de Broglie relation, developed by Louis Victor de Broglie) in the wavelike description of particles are to be found most naturally in Einstein's explanation of the photoelectric effect and in the equations of special relativity. It is an enigma that the seeds of quantum theory lie within one of Einstein's creative explanations of a very puzzling experimental observation, the photoelectric effect. The enigma arises from the fact that Einstein's best-known work, the theory of special relativity, is an entirely separate and remarkably different theory philosophically from that of quantum mechanics.

III. QUANTUM CONCEPTS

A. Early Quantum Theory

Old quantum theory is essentially the patchwork of classical and quantum ideas that were pieced together to yield Planck's theory of black-body radiation, Einstein's explanation of the photoelectric effect, and Bohr's theory of the one-electron atom. Old quantum theory included the concept of the particle nature of radiation (i.e., the photon) and the quantization of the radiation energy in elemental units $h\nu$. However, it did not include the concept of the wave nature of matter.

B. Einstein's Concept of Light as an Energy Quantum

The origin of the word "quantum" in quantum mechanics can be understood from its use in the explanation of some of the properties of light. It was Einstein, in his explanation of the interaction of light with metal surfaces, who gave impetus to the concept that electromagnetic radiation is made up of discrete increments of energy. At that time (1905), light was generally considered to be an electromagnetic wave somewhat similar to water waves or sound waves, with an accompanying energy density that could lead to energy exchange with another body. A light wave in free space may be viewed classically as coupled, time-dependent electric and magnetic fields that are in phase, each varying periodically in time and space. This leads to signal propagation through free space and material media. The electric field of the wave leads to an electric force on charged particles; charged particles can thus be accelerated and thereby gain kinetic energy at the expense of the energy density of the electromagnetic wave. This picture leads to the prediction of electrons in a solid being accelerated to various energies, depending upon the acceleration time, with no sharp cutoff for the maximum energy that can be attained and with no indication that there should be a wavelength- or color-dependence of the effect. It also leads naturally to the conclusion that a more intense light wave of the same wavelength should exert a larger electric force and hence produce more acceleration and consequently higher-energy electrons. In addition, it might be expected that if the light is of extremely low intensity, there would be a measurable time delay before the electron could gain sufficient energy to overcome the surface energy barrier (given by the workfunction) and thereby escape from the metal.

The classical predictions for the interaction of light with solids were not in accord with the experimental data. No time delay was observed. Electron emission from the metal surface was either observed or not observed, depending upon whether the wavelength was less than or greater than some critical wavelength characteristic of the metal used for the experiment. Furthermore, all emitted electrons had a kinetic energy equal to or less than a given maximum energy that increased with the wave frequency. (The wave frequency ν is given by c/λ, where c is the speed of light and λ is the wavelength; it is also the reciprocal of the time period of the oscillation and thus typically has units of cycles per second, termed Hz.)

Einstein pointed out that all of these observations could be explained in a neat way by postulating that the energy transfer between light wave and electron occurs only in well-defined, discrete quantities of size $h\nu$, with h being the fundamental constant introduced by Planck in the year 1900 to resolve the black-body radiation spectrum problem. According to this postulate, light is quantized in elemental units of energy that are incapable of further decomposition. From this viewpoint, the properties of light are similar to the properties of elemental particles, such as the electron and the proton, which have basic units of mass and charge of specific values that are ordinarily incapable of further decomposition. The energy $h\nu$ of each light quantum, called a *photon*, could be transferred to a conduction electron in the metal to enable it to escape from the metal with a kinetic energy given by

$$\mathscr{E}_{\mathrm{K}} = h\nu - \phi, \tag{23}$$

where the workfunction ϕ is the energy step that must be overcome by the electron at the metal surface in order to become free. The action of the photon in promoting electron emission from a metal is illustrated in Fig. 1.

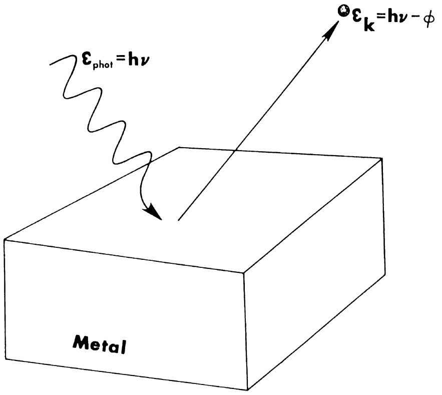

$\mathcal{E}_{\mathrm{phot}} = h\nu$
$\mathcal{E}_{k} = h\nu - \phi$
Metal

FIGURE 1 Photoelectric effect. [Fig. 1.1 in *Quantum Mechanics for Applied Physics and Engineering* by Albert Thomas Fromhold, Jr. (Academic Press, Inc., New York, 1981; Dover Publications, Inc., New York, 1991); reproduced with the permission of Academic Press, Dover Publications, and the author.]

The idea that light has particlelike properties was in one sense a resurrection of the corpuscular theory of light espoused by Newton. Since that theory had long been abandoned, in view of the wavelike properties of diffraction and interference later discovered by Christian Huygens (1629–1695), the particlelike picture invoked by Einstein to explain the photoelectric effect was once again revolutionary. Even Planck, who had chanced upon the explanation of the black-body radiation spectrum by introducing the concept that the radiation in a cavity could only have discrete values for the energy, did not believe that the discreteness was associated with light in any fundamental way. Instead, Plack believed initially that the discreteness proceeded only from constraints on the absorption and emission of radiation by the cavity walls themselves. Light itself generally was viewed as a wave phenomenon, with periodic variations in the electric and magnetic fields but no sharp discontinuities in space and time. To postulate, as Einstein did, that the energy could be localized in space such that it could be exchanged with the electron in a metal instantaneously and in discrete quantities, when the wave nature of light already had been experimentally confirmed, was quite revolutionary.

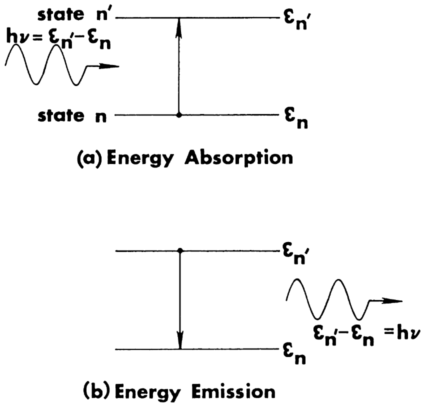
Indeed, it was nearly akin to postulating that light had dual properties.

C. Transitions between Atomic States; Photon Emission

Looking into the phenomenon of light production in a light source (i.e., the origin of optical spectra), one is led to the concept of excited states of the atoms in the source. An atom becomes excited when it absorbs energy in some way (e.g., from the heat associated with a rise in temperature). Such an excited atom can give off energy in the form of a light wave or, more generally speaking, by emitting some form of electromagnetic radiation. From the viewpoint of classical mechanics, the atom can have a continuous range of energies in the excited state and therefore can emit a continuous range of electromagnetic wave energies. [See, for example, Eqs. (4), (7), and (14).] The optical absorption and emission spectra, from this classical viewpoint, would be quite nondescript and, moreover, would vary little from element to element in the periodic table. Such simplicity, however, is belied by the experimental optical spectra. The optical spectrum obtained from a gas discharge is usually very complex, and it is so characteristic of the atoms making up the gas that it can serve as a "fingerprint" identifier of the element making up the gas. From a classical view the most unexpected characteristic of the experimental optical spectrum is the sharpness of peaks in intensity at certain wavelengths (or frequencies). Such peaks are difficult or impossible to explain on the basis of a classical mechanics picture of the atom.

If Einstein's view of light as composed of discrete quantities of energy is to be sustained, then it is natural, if not absolutely necessary, to conclude that these entities are produced by the atoms in the light source. The relationship between light wave frequency and the light quantum energy existent in the photoelectric effect leads one to speculate on the reason for the sharp emission spectrum at certain wavelengths in optical spectra. The spectra, which are so characteristic of the atoms comprising the source, must indicate that the changes in the atom configuration leading to light emission occur in such a way that *individual* quanta of energy $h\nu$ are created and emitted. Figure 2 illustrates the concept of energy absorption and energy emission. The energy $h\nu$ of the photon emitted, for example, is the difference in energy between the initial state $\mathscr{E}_{n'}$ and the energy of the final state $\mathscr{E}_n$ of the quantum system given by

$$h\nu = \mathscr{E}_{n'} - \mathscr{E}_n. \tag{24}$$


state n'
$h\nu = \mathscr{E}_{n'} - \mathscr{E}_n$
state n
$\mathscr{E}_{n'}$
$\mathscr{E}_n$
$\mathscr{E}_{n'}$
$\mathscr{E}_{n'} - \mathscr{E}_n = h\nu$
$\mathscr{E}_n$

(a) Energy Absorption

(b) Energy Emission

FIGURE 2 Quantum transitions involving photon absorption and emission. [Fig. 1.5 in *Quantum Mechanics for Applied Physics and Engineering* by Albert Thomas Fromhold, Jr. (Academic Press, Inc., New York, 1981; Dover Publications, Inc., New York, 1991); reproduced with the permission of Academic Press, Dover Publications, and the author.]

The experimentally unconfirmed expectation on the basis of classical mechanics that the optical spectrum should be rather more or less continuous, with no sudden, sharp changes in intensity with wavelenght, therefore had to be superseded by the view that the changes in the mechanical configuration are catastrophic and instantaneous. Classically, a sudden change in the total energy of an inverse-square force system must be accompanied by a sudden change in the radius of the orbit and also a corresponding change in the rotation frequency of the particle in the orbit. [See Eqs. (4), (7), and (10).] The accompanying change in the energy therefore must occur suddenly, the difference in energy between the two characteristic configurations determining the energy of the emitted photons. The mechanical states of the physical system are thus specific and characterized by fixed energies, with the transition between any two such fixed energy states occurring suddenly with the emission of a fixed quantum of energy. Collapse must occur from one characteristic configuration to another characteristic configuration of lower energy when the photon is created. Each configuration may individually be compared to the intricate workings of a fine watch, with the change in configurations occurring as a sudden transition between two well-defined and self-regulated states.

D. Intensity Peaks in Optical Spectra

The resonance nature of the absorption and emission of electromagnetic radiation by gases found experimentally requires rejection of the simple classical picture of an atom derived from a planetary model of one or more electrons revolving about the nucleus. One is led to the viewpoint that only certain characteristic mechanical configurations of the electrons in an atom are allowed. This provides a plausible explanation of why radiation energy can be abstracted only in units of quanta $h\nu$ of radiation energy. The total energy therefore always changes by discrete photon-energy increments $h\nu$. To the extent that the electron can be viewed as changing speed in its orbital motion to yield the emission of energy, that change in speed must occur in a single-step process. Clearly, an entirely new type of mechanics is required for the description of atomic systems having such properties. Due to the discreteness of the stationary states and the quantum nature of the energy emitted as photons, this new type of mechanics was called quantum mechanics.

E. Ideas of de Broglie, Heisenberg, and Schrödinger

Matter was discovered in 1927 to have wavelike properties by means of the electron diffraction experiments of Clinton Joseph Davisson (1881–1958) and Lester Halbert Germer (1896–1971), and G. P. Thomson (1892–1975). This was preceded (1923) by de Broglie's deduction from special relativity considerations that a particle of energy $h\nu$ and momentum p had a wavelength λ associated with it. In the same way that electromagnetic radiation can be observed to behave in a wavelike manner under certain conditions and in a particlelike manner under other conditions (witness radio-wave interference on the one hand and the photoelectric effect on the other), de Broglie postulated that there was a wave–particle duality for all of nature. Wave–particle duality asserts that nonzero mass particles, such as electrons, protons, neutrons, and atoms, can be observed to behave in a wavelike manner under the proper conditions and in a particlelike manner under different conditions.

This very important concept of particles possessing an intrinsic wave character underlies the Schrödinger equation—that cornerstone of present-day quantum mechanics dating from 1926. The birth of quantum mechanics, in fact, dates to the developments of Werner Karl Heisenberg (1901–1976) in 1925 and of Erwin Schrödinger (1887–1961) in 1926, both of which intrinsically contain the wave properties of matter, although somewhat differently. Whereas Schrödinger directly developed a wave equation to describe the behavior of matter, Heisenberg incorporated the wave properties into a theory in a somewhat different way, setting up a noncommuting matrix operator formulation. Heisenberg was able to show that certain pairs of physical observables represented by noncommuting operators, in principle, could not be measured to arbitrary precision; rather, the more precise the measurement of one member of the pair was, the less precise the knowledge of the other would be. Thus was born the Heisenberg uncertainty principle. This type of uncertainty follows naturally from a wavelike description of matter such as that represented by the Schrödinger equation. The Schrödinger formulation can be used to deduce a matrix formulation of quantum mechanics analogous to Heisenberg's formulation, so in this sense the two theories are equivalent.

F. Bohr Quantized Energy Levels for Hydrogen Atom

It is remarkable that Bohr was able to put together a successful theory of the hydrogen atom, since its formulation in 1913 predated by more than two decades the discovery of the wave diffraction of particles and even predated by a decade the postulate of wave–particle duality by de Broglie and his deduction that a particle of momentum p has a wavelength λ associated with it. Fundamentally, Bohr utilized the data given by experimental optical spectra together with heuristic arguments based on the classical limit. A somewhat different line of reasoning, based on the experimentally confirmed de Broglie relation, is followed here in deducing the quantized energy levels of Bohr.

As a preliminary to the development of the Schrödinger equation utilizing the experimentally verified de Broglie relation

$$\lambda = \frac{h}{p} \tag{25}$$

between the wavelength λ associated with a particle and the momentum p of that particle, where h is Planck's constant having the value 6.6262×10^{-34} Joule-second (J-s), let us use that same relation as a selection device for choosing a series of discrete orbits for an electron imagined to circle a proton from the continuous range of orbits allowed in the purely classical planetary model of the one-electron atom. The key addition is the requirement that a wave associated with a trajectory must be a single-valued function of position on the trajectory. For a circular orbit, this requires that the wavelength be commensurate with the circumference of the orbit, namely

$$\frac{2\pi r}{\lambda} = n, \tag{26}$$

where r is the radius of the circular orbit, λ is the wavelength, and n is any integer ≥ 1. Note that no attempt is made to interpret the meaning of the wave, which is assumed to exist around the circumference of the orbit; instead, only one of the most general properties of waves (i.e., single-valuedness) is relied on when writing the condition given by Eq. (26). The logic of this approach is merely that if the properties of matter are wavelike and if a planetary orbit exists, then the condition given by Eq. (26) should be met. Later in the treatment of the hydrogen atom by means of the Schrödinger equation, the concept of a well-defined planetary orbit will be found to be too naive, except in what is designated the classical limit. This is due to the fact that a wave does not usually have a precise localization. Nevertheless, there do exist well-defined values of the most likely separation distance of the electron relative to the nucleus of the one-electron atom.

Substituting Eq. (25) into (26) gives

$$2\pi pr/h = n \tag{27}$$

which, in terms of the new constant $\hbar$ defined as

$$\hbar = \frac{h}{2\pi}, \tag{28}$$

takes the form

$$pr = n\hbar. \tag{29}$$

Because for a circular orbit, the product pr is the magnitude of the vector angular momentum **L**, Eq. (29) constitutes a restrictive condition on the angular momentum; that is, the values of the angular momentum L are restricted to the discrete set of values L_n given by

$$L_n = n\hbar \qquad (n = 1, 2, 3, \ldots). \tag{30}$$

This is a statement of the quantization of angular momentum. Thus, one arrives at the startling conclusion that not only is angular momentum *conserved* for the one-electron atom, as can be readily deduced from classical mechanics, but also that the new quantum condition establishes that the angular momentum must be *quantized*. The elemental unit for the mechanical angular momentum is thus $\hbar$. Because angular momentum quantization proceeds entirely from the wave description, it is a direct consequence of the wave nature of particles.

Now let us show that the quantization of the angular momentum leads to the quantization of the energy values for the one-electron atom. Introducing the quantum condition given by Eq. (30) into Eqs. (12), (7), and (14) gives discrete values for the radius

$$r_n = \frac{(n\hbar)^2}{KZe^2m} \tag{31}$$

and leads to the following quantized values for the total energy of the Bohr atom

$$\mathscr{E}_{\mathrm{T}} = -\mathscr{E}_{\mathrm{K}} = \frac{-mK^2Z^2e^4}{2n^2\hbar^2} \qquad (n = 1, 2, 3, \ldots). \tag{32}$$

Using $1/(4\pi\varepsilon_0)$ for K, this expression for the total energy can be written as

$$\begin{aligned}\mathscr{E}_n &= \frac{-mZ^2e^4}{32\pi^2\varepsilon_0^2n^2\hbar^2}\\ &= \frac{-mZ^2e^4}{8\varepsilon_0^2n^2h^2} \qquad (n = 1, 2, 3, \ldots).\end{aligned} \tag{33}$$

Thus, the average value of the orbit radius is predicted to increase with an increase in the integer n while the energy increases algebraically from negative values to approach the asymptotic limit of zero corresponding to an unbound state. At the other extreme of small values for n, the negative total energy becomes algebraically smaller, corresponding to tighter binding of the electron and more localization of the electron in the neighborhood of the nucleus. The lowest energy state is given by $n = 1$, so this is the ground state of the one electron atom. Denoting the total energy in this state by $\mathscr{E}_0$ gives

$$\mathscr{E}_0 = \frac{-mZ^2e^4}{32\pi^2\varepsilon_0^2\hbar^2} = \frac{-mZ^2e^4}{8\varepsilon_0^2h^2}. \tag{34}$$

Substituting the values $m = 9.1096 \times 10^{-31}$ kg, $e = 1.6022 \times 10^{-19}$ Coulombs (C), $\hbar = h/2\pi = 1.0546 \times 10^{-34}$ J-s, and $\varepsilon_0 = 8.854 \times 10^{-12}$ F/m gives

$$\begin{aligned}\mathscr{E}_0 &= -Z^2 \times 2.180 \times 10^{-18}\,\mathrm{J}\\ &= -Z^2 \times 13.60\,\mathrm{eV}\end{aligned} \tag{35}$$

as the ground-state energy in electron volts (eV) for the one-electron atom. For hydrogen, $Z = 1$, so the ground-state energy of the hydrogen atom is predicted to be -13.6 eV. The corresponding Bohr radius, as deduced from Eq. (31), is 0.529×10^{-10} m. As will be deduced shortly by using these results to examine the predictions for optical spectra, the theory yields results that are in good agreement with experiment. Thus, this amalgam of classical mechanics and the assumption of a wavelike character for the electron in orbit, as given by the de Broglie relation, leads to a new picture for electrons in atoms.

Needless to say, the electron is too small to be observed directly in its orbit, even if such a picture were tenable from a fundamental standpoint, so the theory cannot be proved or disproved in this way. On the other hand, optical spectra can be measured, so optical measurements can serve as a point of contact between microscopic mechanical models of the atom and the world of observation. The predictions of the model therefore can be tested in this manner. In Section III.H, the Bohr theory will be shown to lead to a reasonably accurate explanation for the optical spectra for the one-electron atom.

Despite the success of the Bohr theory for one-electron atoms, it is incapable of giving realistic predictions for atoms having two or more electrons and, a *fortiori*, fails to give a general explanation of the periodic table for the great variety of elements to be found in nature. Therefore, it can be concluded that a stronger wave theory is needed for understanding and predicting nature. The planetary model, as modified in the simplest possible way by the concept of the wave nature of matter, is sufficiently successful to indicate a promising way to proceed, since it gives some insight into the types of thought processes and concepts required to begin a better treatment.

G. The Heisenberg Uncertainty Relations

There is a still more fundamental problem with the Bohr theory as deduced on the basis of the wave properties of matter inherent in the de Broglie relation. This problem has to do with the basic meaning of the position of an electron on one of the Bohr orbits in the hydrogen atom. The philosophical question is whether or not a physical quantity is actually meaningful if it cannot be measured. The Heisenberg uncertainty principle actually implies that the position of an electron on the smaller orbits in the Bohr model cannot be measured without serious disruption of the electron trajectory itself, as we shall prove later. First of all, let us examine one approach that can be used to rationalize the uncertainty relation.

Let us consider the experimental aspects of a position measurement of the location of a point mass by means of a microscope utilizing electromagnetic waves of wavelength λ_{photon} and frequency $\nu_{\text{photon}} = c/\lambda_{\text{photon}}$. It is a well-known fact that the resolution of a microscope is limited by the wavelength of the light utilized for the measurement such that the uncertainty Δx_{mass} of the position measurement will be of the order of (or greater than) the wavelength:

$$\Delta x_{\text{mass}} \geq \lambda_{\text{photon}}. \tag{36}$$

However, as already discussed, photons have momentum $p = h/\lambda$, and the conservation of momentum when the measuring photon scatters off the point mass will cause an uncertainty in the final momentum of the point mass of this order—namely, $\Delta p_{\text{mass}} \approx h/\lambda_{\text{photon}}$. Therefore, it must be concluded that, following the measurement, $\Delta x_{\text{mass}} \Delta p_{\text{mass}} \geq (\lambda_{\text{photon}})(h/\lambda_{\text{photon}}) = h$. Thus, there is some lower limit to the precision with which the two physical variables of momentum and position of a particle can be measured. This conclusion is consistent with a rigorous version of what is known as the position—momentum form of the Heisenberg uncertainty relation, which states that the *minimum* value of $(\Delta x)(\Delta p)$ is of the order of $\frac{1}{2}\hbar$. There is no particular limit to the maximum value. Therefore,

$$\Delta x \Delta p \geq \frac{1}{2}\hbar. \tag{37}$$

Two variables for which the uncertainty relation holds are known as complementary variables.

An alternate pair of complementary variables is given by the energy of a particle in a quantum state and the lifetime of the particle in the state. This means that there is also a lower limit to the product $\Delta\mathscr{E}\Delta t$, where $\Delta\mathscr{E}$ is the uncertainty of the energy of the particle and Δt is the uncertainty in the time the particle will remain in that state. Thus, in analogy with Eq. (37), $\Delta\mathscr{E}\Delta t \geq \frac{1}{2}\hbar$.

For a meaningful electron orbit, the uncertainty in position of the electron must be less than the radius of the orbit. Applying Eq. (37) to this situation yields

$$\Delta p_n > \frac{\hbar}{2r_n} \tag{38}$$

for a Bohr orbit of radius r_{n}. The radius r_n is given by Eq. (31), so that for $Z = 1$, $r_n = 4\pi\varepsilon_0 n^2\hbar^2/me^2$. Thus

$$\Delta p_n > \frac{me^2}{8\pi\varepsilon_0 n^2\hbar}. \tag{39}$$

The corresponding total energy $\mathscr{E}_{\text{T}}$ given by Eq. (32) is $\mathscr{E}_n = -me^4/32\pi^2\varepsilon_0^2 n^2\hbar^2$. The momentum p_n for state n obtained by means of the general relation $p = (2m\mathscr{E}_k)^{1/2}$ is given by $p_n = (2m|\mathscr{E}_n|)^{1/2}$, since $\mathscr{E}_{\text{K}} = -\mathscr{E}_{\text{T}}$ according to Eq. (32). Substituting the expression for $\mathscr{E}_n$ then gives $p_n = me^2/4\pi\varepsilon_0 n\hbar$. Now let us establish the requirement on $\Delta\mathscr{E}_n$ provided by Δp_n evaluated above. Again using $\mathscr{E}_{\text{K}} = -\mathscr{E}_{\text{T}}$, with $\mathscr{E}_{\text{K}} = p^2/2m$, it is found that $\Delta\mathscr{E}_{\text{K}} = (p/m)\Delta p$, so that $\Delta\mathscr{E}_n = (p_n/m)\Delta p_n$. Substituting the above expressions for p_n and Δp_n then gives

$$\Delta\mathscr{E}_n > \frac{me^4}{32\pi^2\varepsilon_0^2 n^3\hbar^2} = |\mathscr{E}_n|/n. \tag{40}$$

When $n = 1$, the ground-state Bohr orbit having radius r_1 is denoted by a and the ground-state energy $\mathscr{E}_1$ is denoted by $\mathscr{E}_0$. For this case, $\Delta\mathscr{E}_0 > |\mathscr{E}_0|$. This surprising result indicates that the uncertainty introduced into the energy by an attempted measurement of the position of the electron on its ground-state orbit exceeds the binding energy itself, so the stable configuration will be seriously disrupted by the measurement, if not totally destroyed. Thus, if one subscribes to the debatable tenet that a quantity is not physically meaningful unless it can be experimentally measured, then the very meaning of an electron orbit associated with the ground-state energy of the Bohr model is brought into serious question. In any event, an electron cannot be observed directly in orbit about a proton, so it is impossible to say exactly what the separation distance is between electron and proton at any given instant. However, the energy absorption and emission due to transitions of the hydrogen atom between various states of excitation can be measured. This is indeed an experimental point of contact with the theory. Therefore, let us examine the predictions of the Bohr theory for the optical spectrum of hydrogen.

H. Optical Spectrum of Hydrogen

If the electron in the hydrogen atom with energy corresponding to the integer n suddenly undergoes a transition to an energy corresponding to a different integer n', with the energy difference being positive so as to create a photon of energy $h\nu$, then energy conservation gives the relation

$$h\nu = \mathscr{E}_n - \mathscr{E}_{n'} = -\mathscr{E}_0\left[\left(\frac{1}{n'}\right)^2 - \left(\frac{1}{n}\right)^2\right], \qquad (41)$$

where $\mathscr{E}_0$ is given by Eq. (34), with $Z = 1$. Since for the photon

$$h\nu = \frac{hc}{\lambda} = 2\pi\hbar\frac{c}{\lambda} \qquad (42)$$

the relation given by Eq. (41) can be used to write

$$\frac{1}{\lambda} = \frac{h\nu}{2\pi\hbar c} = \frac{-\mathscr{E}_0}{2\pi\hbar c}\left[\left(\frac{1}{n'}\right)^2 - \left(\frac{1}{n}\right)^2\right]. \qquad (43)$$

This is usually written in the form

$$\frac{1}{\lambda} = R\left[\left(\frac{1}{n'}\right)^2 - \left(\frac{1}{n}\right)^2\right], \qquad (44)$$

where

$$R = \frac{-\mathscr{E}_0}{2\pi\hbar c} = \frac{me^4}{64\pi^3\hbar^3\varepsilon_0^2 c} \qquad (45)$$

is known as the Rydberg constant. Thus, transitions from the various excited states ($n > 1$) to the ground state ($n' = 1$) leads to a series of spectral lines with frequencies $\nu = c/\lambda$ given by

$$\nu_{n\to 1} = cR\left[1 - \left(\frac{1}{n}\right)^2\right] \qquad (n = 2, 3, 4, \ldots). \qquad (46)$$

This series of lines is known as the Lyman series, which is in the ultraviolet region of the spectrum. Yet a second series of spectral frequencies can be generated by transitions from excited states with $n > 2$ to the state $n' = 2$. It can be noted that these frequencies are given by

$$\nu_{n\to 2} = cR\left[\left(\frac{1}{2}\right)^2 - \left(\frac{1}{n}\right)^2\right] \quad (n = 3, 4, 5, \ldots). \qquad (47)$$

The spectral lines in this series, known as the Balmer series, have wavelengths in the near ultraviolet and visible region of the spectrum. Similar series of lines determined from $n' = 3, 4,$ and 5 can be written using the above prescription; these three series, labeled Paschen, Brackett, and Pfund, respectively, have spectral lines with wavelengths in the infrared region of the spectrum.

These various series of spectral lines were known from experimental observation long before the invention of the Bohr model of the atom and indeed constituted some of the major evidence that classical mechanics was inadequate for the treatment of atomic systems. The frequency relationships deduced from the Bohr theory agree quite well with the experimental spectra. Even better agreement is obtained when the finite mass of the nucleus is included in the theory; the correction comes about because electron and proton revolve about the center of mass, giving a slightly smaller orbit radius than the one obtained above (assuming the radius is the same as the separation distance between electron and proton). This is the so-called reduced mass effect, which is purely classical in nature.

I. The Laser

The absorption and emission of light due to electron transitions between quantized energy levels provide the basis for one of the most powerful research tools developed over the past 50 years—namely, the laser. The key property of a laser beam is its phase coherence. The phase coherence of the beam is to be contrasted with the random phase relationships between the light emitted from various regions of an ordinary light source, such as an incandescent filament. The phase-coherent beam in a laser is produced by reflecting the emitted light back and forth between parallel mirror surfaces bounding the emitting medium, so that any light already emitted triggers additional emission and influences the phase of such additional light emission from the excited atoms of the medium. The phase of the newly emitted light turns out to be the same as that of the already emitted light. It is not intuitive from a quantum mechanical viewpoint that this should occur, but it is as one might expect on the basis of classical physics, since the electric field provides the accelerating force for the atomic electrons. That phase coherence does occur is, in fact, basic to the nature of the process referred to as stimulated emission. Stimulated emission is to be distinguished from spontaneous emission, which prevails in ordinary light sources which lack phase coherence.

The emission of light in a laser can initiate due to prior electronic transitions in atoms in a gas (e.g., a helium—neon mixture) or in a solid (e.g., ruby) that have been preexcited by some process, such as the flash of an ordinary discharge lamp. As the phase-coherent beam builds in intensity within the laser, a portion of it is allowed to emerge continuously from the active lasing medium by a partial transmission through one of the mirrored surfaces.

The intense phase-coherent beams produced by lasers allow many enhanced experiments involving light interference to be carried out. The laser has enabled the speed of light to be measured to much greater precision than ever before, such that the wavelength of a laser beam can be used as an accurate standard for the measurement of length. The angular divergence of a laser beam can be held to quite small values, so that reflectance of a measurable signal from very great distances (e.g., from earth to moon and return) becomes possible. By measuring the time for round-trip light travel, large distances can be measured more accurately than was posible previously. The Michaelson–Morley experiment for determining the speed of light in different directions relative to the earth's instantaneous velocity vector in its orbit about the sun also can be performed with greater precision with the aid of the laser.

IV. WAVE MOTION

A. Development of the Wave Equation for Transverse Vibrations

The essential equations for classical wave motion constitute the foundation for developing intuitive insight into the fundamentals of quantum mechanical wave equations for particles. It is well known that wave motion can describe as varied phenomena as the transverse displacement of a wire in tension, the propagation of sound in gases, the motion of electromagnetic radiation in free space, and the longitudinal displacements associated with elastic waves in a solid. All of these classical phenomena can be treated theoretically by means of the same differential equation technique, which moreover proves to be adequate for describing the wave properties of particles.

Consider, for example, the basic classical problem of the time and position dependence of the transverse displacement of a vibrating wire, string, or rope, as indicated in Fig. 3. The transverse displacement $y(x, t)$ of a thin wire under tension T, with mass per unit length along the x-direction given by ρ, is determined as a function of time t by a differential equation that can be obtained from a straightforward application of Newton's second law of motion, $F = ma$, where F is the force and a is the acceleration. A net force F_y acting in the y-direction on a length Δx produces an acceleration a_y that is inversely proportional to the mass $\rho\Delta x$,

$$F_y = (\rho\Delta x)a_y = (\rho\Delta x)\left(\frac{d^2y}{dt^2}\right). \tag{48}$$

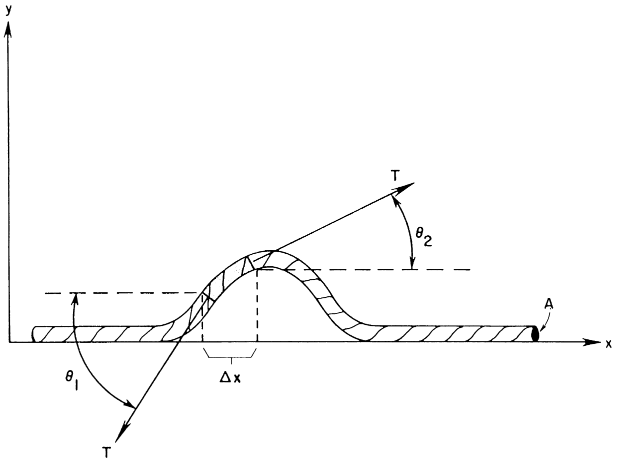


FIGURE 3 Transverse wave.

The net force F_y is given by the difference between the y-directed components of the uniform axial tension force T at the two ends of the element Δx. Considering θ to measure the angle in radians between the wire and the horizontal line representing the wire when it has no transverse displacement, θ can be visualized as varying with position x along the wire and with time t as the wire vibrates. The tensile force T is considered to have a line of action that is always parallel to the wire. This leads to a projected component in the y-direction given by $(T \sin\theta)$. At the ends of the segment Δx the tensile force T acts in opposite directions, as is required of a tensile force. If the segment is not to be accelerated in the x-direction, then the x-components of these end forces must essentially cancel each other. The component of T projected in the x-direction is given at any point by $(T \cos\theta)$, so this cancellation condition is met adequately if θ is small enough. In the small θ limit, the net force in the y-direction on the segment Δx, namely,

$$F_y = T \sin\theta_{x+\Delta x} - T \sin\theta_x \tag{49}$$

reduces to the approximation

$$F_y \simeq T\theta_{x+\Delta x} - T\theta_x \tag{50}$$

which, in turn, can be approximated by

$$F_y = T \tan\theta_{x+\Delta x} - T \tan\theta_x. \tag{51}$$

Because $\tan\theta$ is the slope $\partial y/\partial x$, this last equation is equivalent to

$$F_y = T\left[\left(\frac{\partial y}{\partial x}\right)_{x+\Delta x} - \left(\frac{\partial y}{\partial x}\right)_x\right] \tag{52}$$

Equating this expression for F_y to the partialderivative equivalent of Eq. (48) gives

$$(\rho\Delta x)\left(\frac{\partial^2 y}{\partial t^2}\right) = T\left[\left(\frac{\partial y}{\partial x}\right)_{x+\Delta x} - \left(\frac{\partial y}{\partial x}\right)_x\right] \tag{53}$$

Dividing through by Δx and taking the limit $\Delta x \to 0$ gives

$$\rho\frac{\partial^2 y}{\partial t^2} = T\frac{\partial^2 y}{\partial x^2} \tag{54}$$

The unit of tension T in SI units is Nt and the units of ρ are kg/m, so T/ρ has the dimensions of m^2/s^2, the square of a speed. Denoting the square root of this quantity by

$$v_p = \left(\frac{T}{\rho}\right)^{1/2} \tag{55}$$

where v_p has the units of velocity, the equation for a vibrating wire takes the form

$$\frac{\partial^2 y}{\partial x^2} = \frac{1}{v_p^2}\left(\frac{\partial^2 y}{\partial t^2}\right) \tag{56}$$

This is the well-known classical wave equation. It is a linear, second-order differential equation that governs, in the present example, the transverse displacement $y(x, t)$ as a function of position along the wire as time proceeds. Both standing-wave modes and running-wave modes are possible, depending upon the boundary conditions imposed on the wire.

The classical wave equation given by Eq. (56) also has a three-dimensional form to describe motion in arbitrary directions in space. There are various ways to generalize Eq. (56), but it is simplest first of all to think of wave motion along the z-axis, in contrast to the x-axis. The simple change in dependent variable from x to z naturally yields the relevant equation. It is clear that to include the three independent directions in space will require three terms involving spatial derivatives in place of the single term in Eq. (56). If these three Cartesian coordinates are denoted by x, y, and z, then it is necessary to use some alternate notation for the wave displacement. Denoting the wave displacement by $\psi(x, y, z, t)$, the classical three-dimensional wave equation then can be written as

$$\nabla^2\psi = \left(\frac{1}{v_p}\right)^2 \left(\frac{\partial^2\psi}{\partial t^2}\right), \tag{57}$$

where ∇^2 symbolizes the differential operator $(\partial^2/\partial x^2) + (\partial^2/\partial y^2) + (\partial^2/\partial z^2)$.

B. Solutions to the Classical Wave Equation

Let us attempt a trial solution for Eq. (56) of the form

$$y = C\exp[i(kx - \omega t)], \tag{58}$$

where, at the moment, C, k, and ω are unspecified constants, perhaps even complex numbers. Substituting Eq. (58) into Eq. (56) gives a condition on the constants ω and k in terms of v_p:

$$k^2 = \frac{\omega^2}{v_p^2}. \tag{59}$$

The complex form of the trial solution is troublesome for those who are more concerned with physical phenomena than with mathematics, since the displacement is a real quantity. In point of fact, the mathematical solution, in itself, does not have any physical interpretation associated with it. To be able to clothe the mathematical solution with physical content, it is often necessary to restrict somewhat the range of solutions that can be accepted.

The physically meaningful solutions for the present problem of a vibrating wire are thus given by

$$[y]_{\substack{\text{physically}\\ \text{meaningful}\\ \text{subject}}} = \mathcal{R}e\{C\exp[i(kx - \omega t)]\}. \tag{60}$$

where $\mathcal{R}e$ means "take the real part of." It is expedient at this point to restrict the constants k and ω to real values. However, it proves very useful, as will be shown, to maintain the constant C in its most general complex form,

$$C = D\exp(i\Theta), \tag{61}$$

where D and Θ are real numbers. The trial solution therefore can be written in the form

$$y = D\exp[i(kx - \omega t + \Theta)], \tag{62}$$

thus giving real solutions of the form

$$y = D\cos(kx - \omega t + \Theta). \tag{63}$$

The right-hand side of Eq. (63) has the sort of space dependence and time dependence seen for vibrating wires in the laboratory. Larger values of ω yield shorter repetition periods in time, and larger values of k yield shorter repetition periods in space. Considering that a cosine function repeats itself when its phase is increased by 2π, it can be concluded that $k\lambda = 2\pi$ and $\omega r = 2\pi$ give the basic spatial unit λ and temporal unit τ for repetition. These parameters are called *wavelength* and *period*, respectively. Note that

$$k = \frac{2\pi}{\lambda} \tag{64}$$

and

$$\omega = \frac{2\pi}{\tau}. \tag{65}$$

The temporal frequency ν is the reciprocal of the period τ, or

$$\nu = \frac{1}{\tau}, \tag{66}$$

so that

$$\omega = 2\pi\nu. \tag{67}$$

Here, the explicit assumption was made that the constants ω and k are real. For wave motion, this is consistent with the physical interpretations relating these parameters to the temporal frequency and the reciprocal of the spatial periodicity.

C. Phase Velocity of Waves

It is a universal property of wave motion that

$$\frac{\omega}{k} = \frac{2\pi\nu}{2\pi/\lambda} = \lambda\nu = \frac{\lambda}{\tau}, \tag{68}$$

and this ratio gives the speed v_{phase} of the sinusoidal wave, where τ is the period for one temporal oscillation. This is readily understood by visualizing a moving sinusoidal spatial wave pass by a given point in space, the length λ

passing in time τ. The speed of an individual sinusoidal wave is called *phase velocity*. and is denoted by

$$v_{\text{phase}} = \frac{\pm\omega}{k}. \tag{69}$$

Therefore, Eq. (55) states that the quantity $v_p = (T/\rho)^{1/2}$ is the speed with which the sinusoidal wave moves along the wire. The phase velocity may depend upon frequency or wavelength in some instances. Note that in the present example, however, the phase velocity depends directly upon the square root of the tension in the wire and inversely upon the square root of the mass per unit length of the wire, independent of the frequency or the wavelength of the wave.

D. Development of the Wave Equation for Electromagnetic Waves

It can be readily shown that electromagnetic waves in free space also satisfy a wave equation analogous to Eq. (57) for transverse waves along a wire and thus have solutions that are mathematically the same. To develop this wave equation, consider Maxwell's equations of electromagnetic theory

$$\nabla \cdot \mathbf{D} = \rho \tag{70}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{71}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{72}$$

$$\nabla \times \mathbf{H} = \mathcal{J} + \frac{\partial \mathbf{D}}{\partial t}, \tag{73}$$

where $\mathbf{D}$ is the electric displacement vector, $\mathbf{B}$ is the magnetic induction vector, $\mathbf{E}$ is the electric field vector, $\mathbf{H}$ is the magnetic field vector, $\mathcal{J}$ is the electric current density vector, and ρ is the electric charge density. In free space, the linear relations

$$\mathbf{D} = \varepsilon_0 \mathbf{E} \tag{74}$$

$$\mathbf{B} = \mu_0 \mathbf{H} \tag{75}$$

are applicable, where ε_0 is the electric permittivity of free space and μ_0 is the magnetic permeability of free space. In the absence of free charge and with no electric current density, the following relations are obtained:

$$\nabla \cdot \mathbf{E} = 0 \tag{76}$$

$$\nabla \cdot \mathbf{H} = 0 \tag{77}$$

$$\nabla \times \mathbf{E} = -\mu_0 \frac{\partial \mathbf{H}}{\partial t} \tag{78}$$

$$\nabla \times \mathbf{H} = \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t}. \tag{79}$$

Taking the vector curl of Eqs. (78) and (79) yields

$$\nabla \times \nabla \times \mathbf{E} = -\mu_0 \frac{\partial}{\partial t}(\nabla \times \mathbf{H}) \tag{80}$$

$$\nabla \times \nabla \times \mathbf{H} = \varepsilon_0 \frac{\partial}{\partial t}(\nabla \times \mathbf{E}). \tag{81}$$

Applying the vector identity

$$\nabla \times \nabla \times \mathbf{V} = \nabla(\nabla \cdot \mathbf{V}) - \nabla^2 \mathbf{V}, \tag{82}$$

which is valid for any vector $\mathbf{V}$, to the left-hand sides and simultaneously utilizing the relations $\nabla \cdot \mathbf{E} = 0$ and $\nabla \cdot \mathbf{H} = 0$ given by Eqs. (76) and (77) yields the following relations:

$$\nabla^2 \mathbf{E} = \mu_0 \frac{\partial}{\partial t}(\nabla \times \mathbf{H}) \tag{83}$$

$$\nabla^2 \mathbf{H} = -\varepsilon_0 \frac{\partial}{\partial t}(\nabla \times \mathbf{E}). \tag{84}$$

Substituting the expressions for $\nabla \times \mathbf{E}$ and $\nabla \times \mathbf{H}$ then reduces Eqs. (83) and (84)

$$\nabla^2 \mathbf{E} = \mu_0 \varepsilon_0 \frac{\partial^2 \mathbf{E}}{\partial t^2} \tag{85}$$

$$\nabla^2 \mathbf{H} = \mu_0 \varepsilon_0 \frac{\partial^2 \mathbf{H}}{\partial t^2}. \tag{86}$$

These are vector equations; each represents three scalar equations, one for each of the three vector components. By considering any one component of these fields (e.g., E_x, E_y, E_z, H_x, H_y, or H_z) and calling the particular component ψ, Eqs. (85) and (86) lead to the three-dimensional wave equation

$$\nabla^2 \psi = \left(\frac{1}{c^2}\right)\left(\frac{\partial^2 \psi}{\partial t^2}\right) \tag{87}$$

where, in the present problem, the phase velocity c is determined from the physical constants μ_0 and ε_0 in accordance with $c = (\mu_0 \varepsilon_0)^{-1/2}$. This is the propagation velocity for electromagnetic waves in free space. Substituting the values $\varepsilon_0 = 8.854 \times 10^{-12}$ F/m and $\mu_0 = 4\pi \times 10^{-7}$ henry/m (H/m) gives the speed of light in free space as $c = 2.998 \times 10^8$ m/s.

Equation (87) has particularly simple solutions depending on position $\mathbf{r}$ and time t. For example, plane waves represented by $A \cos(\mathbf{k} \cdot \mathbf{r} - \omega t + \alpha)$ satisfy the equation, where A, $\mathbf{k}$, ω, and α are constants. The magnitude of the vector $\mathbf{k}$ determines the wavelength λ, so that

$$\lambda = \frac{2\pi}{|\mathbf{k}|} = 2\pi \big/ \left(k_x^2 + k_y^2 + k_z^2\right)^{1/2}. \tag{88}$$

The wave is traveling at phase velocity $c = \omega/|\mathbf{k}|$ in the direction of $\mathbf{k}$. The amplitude A of the wave is the maximum value of the transverse field component represented by ψ.

E. Dispersion Relations for Waves

The ω versus k relation for any wave is called the dispersion relation for that wave. The dispersion relation always gives the magnitude of the phase velocity v_{phase} of the wave in accordance with $v_{\text{phase}} = \omega/k$. With this generalization, Eq. (57) can be viewed as a more general equation for one-dimensional classical wave motion. The same form of the equation holds for electromagnetic wave propagation in free space, according to Eq. (87). The phase velocity is then the speed of light, and the appropriate dispersion relation is

$$\omega = ck. \tag{89}$$

Light propagates in a dielectric medium according to essentially the same wave equation as that for free space. In that case, however, the phase velocity depends upon the refractive index n of the medium, so that

$$v_{\text{phase}} = \frac{c}{n}, \tag{90}$$

where n usually varies with the wavelength. The action of a prism in dispersing the colors from incident white light, for example, is primarily due to the fact that the phase velocity is color dependent. Red light travels approximately 0.5% faster than blue light in fused quartz. Both colors travel in fused quartz at approximately two-thirds the speed of light in free space, but since the speed of blue light is decreased slightly more than the speed of red light as the white light enters the quartz from the vacuum, the blue light is bent more toward the normal direction than the red light. The colors thus become dispersed as the white light enters the prism at an angle with respect to the normal to the surface of the prism. The dispersion relation for this situation is

$$\omega = v_{\text{phase}}k = \frac{ck}{n(k)}, \tag{91}$$

where n explicitly depends on k, denoting a wavelength-dependence since $\lambda = 2\pi/k$. This example also will be useful in understanding group velocity in the development that follows.

F. Boundary Conditions

The next consideration is the type of boundary conditions that are imposed by the physical situation. If the endpoints of a wire of length L are fixed at positions $x = 0$ and $x = L$, then the fact that the displacements must be zero at these points leads to the requirement that the wave must have stationary nodes at these points. A similar type of boundary condition holds for electromagnetic waves trapped in a highly conducting metal box, since the metal walls cannot support an electric field. Equation (63) cannot satisfy such boundary conditions as time progresses, but two such solutions having the same magnitude k but differing in sign of k, with properly chosen phases, can be superimposed to yield a solution that can satisfy the condition. For example, the superposition solution $y = D[\sin(kx - \omega t) + \sin(kx + \omega t)] = 2D \sin(kx) \cos(\omega t)$ satisfies the boundary condition at $x = 0$; the boundary condition at $x = L$ is satisfied if the wavelength is chosen to have any value $\lambda = L/n$, where n can be any positive integer. The amplitude of this wave is $2D$. The nodes ($y = 0$) and extremum values ($y = \pm 2D$) of the wave do not change position x in time t, so this solution is called a *standing wave*.

On the other hand, for an infinitely long wire (or an unbounded medium for electromagnetic wave propagation), there may be no constraints on the displacement at any particular position, so the solution described by Eq. (63) may be quite acceptable. The particular values of position x for which the wave has nodes and extremum values change with time t. This wave is not stationary and thus is called a *running wave*. For positive values of ω and k, Eq. (63) shows that a given value of the phase is maintained as time increases if one focuses on an observation point x' that moves along the x-axis linearly with t. Thus, the phase velocity ω/k is said to be constant and the wave moves in the positive x-direction. If k is negative but ω is again chosen to be positive, as is conventional, then the phase velocity is negative and the wave moves in the negative x-direction. The standing wave constructed to satisfy the fixed-boundary conditions situation can be viewed simply as the superposition of two equal-amplitude running waves having the same frequency and wavelength but moving in opposite directions. Considerations of more complicated superpositions follow.

G. Superposition Solutions

An arbitrary superposition of solutions to the linear, time-dependent wave equation considered above also satisfies the same equation. There is no way for the terms to mix as products if the equation is linear, so the terms for each solution can individually add to zero. The superposition could be written as a sum of terms having different weighting coefficients D_j, so that

$$y(x, t) = \sum_j D_j \cos(k_j x - \omega_j t + \Theta_j), \tag{92}$$

where each ratio ω_j/k_j has the appropriate value of v_{phase}. The superposition could yield a shape for $y(x)$ at a given time t that differs markedly from any of the sinusoidal components. If each component moves in the same direction at the same speed, corresponding to a value of v_p, which is a fixed constant, then motion would be anticipated but not distortion of shape. In wave packets considered below, it may happen that different component waves have different phase velocities. That is, the phase velocity may be a function of the frequency (or wavelength). Variations in the phase velocity can lead to a change in shape, since phase relationships among the individual components continually change if they are traveling at different speeds.

H. Group Velocity of Waves

It is known from the theory of Fourier series and Fourier integrals that the superposition of sinusoidal waves can yield almost any physically reasonable periodic or nonperiodic function. Thus, by superposing solutions of different wavelengths, nearly any function shape can be generated at any given time. The superposition of waves bearing a harmonic relationship to one another yields a spatially periodic function, whereas the superposition of waves having amplitudes over a continuous narrow band of frequencies yields a spatially localized function. The waves in the superposition may or may not have the same phase velocity. Consider, for example, the case of electromagnetic wave motion in dielectrics, where the physical properties of the medium determining the wave speed are frequency- or wavelength-dependent. Whether or not the shape changes over time, it will generally move. The problem to address now is the evaluation of the speed of motion of the shape.

To simplify the problem, let us confine our attention to a narrow band of wavelengths, corresponding to a narrow range of k values centered about some central value k_0. For example, a pulse of light having a certain shape could have a range of wavelengths closely centered about the green 5461 Å line of mercury. If the amplitude is a continuous function of the wavelength over the band, in contrast to the superposition of a finite number of discrete wavelength components, then an amplitude function $\chi(k)$ can be defined that peaks at k_0 and has very low values outside the narrow band of interest. It is known from the theory of Fourier integrals that the shape of $\chi(k)$ as a function of k depends directly upon the position-dependent shape of the pulse. If $\psi(x, 0)$ denotes the pulse in space at time $t = 0$, then a Fourier inversion directly yields $\chi(k)$. It can be shown that the widths of the two functions are related inversely: the broader $\chi(k)$ is, the more narrow $\psi(x, 0)$ will be, and vice versa. This is a manifestation of a relationship between the spread (or uncertainty) in wavelength versus the spread (or uncertainty) in position of the wave packet. This point is crucial for the Heisenberg uncertainty principle but is not directly relevant to the present development for pulse velocity.

Considering any reasonable shape for $\chi(k)$, the superposition can be written

$$\psi(x, t) = \int_{-\infty}^{\infty} \chi(k) \exp[i(kx - \omega t)]\, dk. \tag{93}$$

This form has the character of a complex Fourier integral. Alternately, the same problem could be treated in terms of integrals involving the real functions, sine and cosine, with the real argument $(kx - \omega t)$ replacing the imaginary argument $[i(kx - \omega t)]$ of the exponential form. To obtain general properties of the superposition, no specific choice is made regarding the functional form of $\chi(k)$, except that it be chosen to be moderate in functional behavior and smooth. In general, it is complex, with the roles of the real and imaginary parts being exactly the same as the roles of the real and imaginary parts of $C = D\exp(i\Theta)$ utilized in Section B—namely, to supply both amplitude and phase information. In this situation, both are supplied as a function of k. Let us assume that the dispersion relation ω versus k is known for the wave in question in the neighborhood of k_0, where the components are presumed to have larger amplitudes. These conditions are then sufficient to use a Taylor series expansion of $\chi(k)$ on k about the value k_0:

$$\omega(k) = \omega(k_0) + \left[\frac{d\omega(k)}{dk}\right]_{k=k_0}(k - k_0) + \frac{1}{2!}\left[\frac{d^2\omega(k)}{dk^2}\right]_{k=k_0}(k - k_0)^2 + \cdots. \tag{94}$$

Substituting the first two terms of this expansion yields the following result for the wave packet under consideration:

$$\psi(x, t) = \int_{-\infty}^{\infty} \chi(k) \exp\left[i\left(kx - \left\{\omega(k_0) + \left[\frac{d\omega(k)}{dk}\right]_{k=k_0}(k - k_0)\right\}t\right)\right] dk. \tag{95}$$

In terms of the defined quantities

$$v_{\text{group}} = \left[\frac{d\omega(k)}{dk}\right]_{k=k_0} \tag{96}$$

$$\beta_0 = \omega(k_0) - k_0 v_{\text{group}} \tag{97}$$

Eq. (95) becomes

$$\psi(x, t) = \int_{-\infty}^{\infty} \chi(k) \exp\{i[k(x - v_{\text{group}}t) - \beta_0 t]\}\, dk. \tag{98}$$

To find out where ψ is peaked in space at any given t, one can consider the real and imaginary parts of this expression individually. Writing

$$\chi(k) = \chi_r(k) + i\chi_i(k), \tag{99}$$

then the real part of Eq. (98) is

$$\psi(x,t)\Big|_{\substack{\text{real}\\ \text{part}}} = \int_{-\infty}^{\infty} \chi_r(k)\cos[k(x - v_{\text{group}}t) - \beta_0 t]\,dk - \int_{-\infty}^{\infty} \chi_i(k)\sin[k(x - v_{\text{group}}t) - \beta_0 t]\,dk. \tag{100}$$

The imaginary part can be obtained similarly. Focusing first on the cosine integral, the coefficient of k (namely, $x - v_{\text{group}}t$), will determine how rapidly the cosine function oscillates as k is varied during the integration over k. Rapid oscillations lead to much cancellation between positive and negative contributions to the integral, with the consequence that the integral will not be large. On the other hand, very small values of the coefficient mean far fewer oscillations over the band of wavelengths where $\chi_r(k)$ is significant, and the smaller amount of cancellation means that the integral will be larger. The same arguments can be employed for the sine function integral. Thus, for a given t, the maximum value of ψ occurs at the position where the coefficient is zero—namely, at $x = v_{\text{group}}t$. Since this position of the peak moves with the velocity v_{group}, the reason for calling this quantity the group velocity is apparent. The identical state of affairs holds for the imaginary part of ψ. Thus, a very general expression for the one-dimensional group velocity is given by Eq. (96).

The additional term $\beta_0 t$ in Eq. (100) can be written as a phase factor $\Theta(t)$—namely, $\Theta(t) = \beta_0 t$—which increases linearly with the time. Because it does not depend upon the variable k of integration or the position x, it is not a very important quantity for determining how relatively large ψ will be as a function of x. Its primary effect is to shift the phase of all superimposed waves by the same constant factor Θ at any given time t, which leads to a modulation of the spatial function over time.

More effects come into play if the third term in the Taylor series expansion of $\omega(k)$ given in Eq. (94) is brought into the wave-packet integral. The packet can then spread and possibly change shape as time evolves.

Let us now apply our results for the group velocity to the situation of electromagnetic wave propagation through a dielectric. Presuming wavelengths in a narrow band to be superimposed to form the packet, the group velocity can be obtained in terms of the refractive index $n(k)$ of the dielectric. Consider a narrow band of wavelengths of green light centered about the 5461 Å line of mercury as those waves travel in fused quartz. The phase velocity being $c/n(k)$, the dispersion relation is $\omega(k) = ck/n(k)$, so the group velocity is

$$\begin{aligned} v_{\text{group}} &= \left[\frac{d\omega(k)}{dk}\right]_{k=k_0} \\ &= \left\{\frac{c}{n(k)}\left[1 - \frac{k}{n(k)}\frac{dn(k)}{dk}\right]\right\}_{k=k_0}. \end{aligned} \tag{101}$$

Since $k = 2\pi/\lambda$, it follows that

$$\frac{dn}{dk} = \frac{dn/d\lambda}{dk/d\lambda} = -\left(\frac{\lambda^2}{2\pi}\right)\left(\frac{dn}{d\lambda}\right). \tag{102}$$

Thus, an alternate form of the group velocity for this situation is

$$v_{\text{group}} = \left\{v_{\text{phase}}\left[1 + \frac{\lambda}{n(k)}\frac{dn(k)}{d\lambda}\right]\right\}_{k=k_0}. \tag{103}$$

Consider $n(k)$ in this expression in terms of its λ-dependence. The phase velocity v_{phase} is that for the central component wave in the packet—namely, at wavelength 5461 Å. Note from this result that for situations in which the refractive index is independent of λ, the group velocity is equal to the phase velocity. Estimates for fused quartz, however, give $n = 1.46$ and $dn/d\lambda = -4 \times 10^{-6}$ Å^{-1} at $\lambda = 5461$ Å. Using these numbers in Eq. (103) gives $v_{\text{group}}/v_{\text{phase}} = 1 - 0.015 = 0.985$. Thus, the group velocity for the wave packet of green light in fused quartz is $\sim$1.5% less than the phase velocity. This may seem to be an unimpressive figure until it is realized that the similarly small difference in phase velocities between 4500 Å and 6500 Å yields the dispersion in a fused-quartz prism that enables the splitting of incident white light into its spectrum of colors.

Although it is interesting to consider that the group and phase velocities can differ, as shown here, there is actually nothing so mysterious about it. The individual sinusoidal waves extend throughout space and, in this sense, are nonlocalized. Any given point on any one of the waves, such as one of the extremum points or one of the nodes, moves with the phase velocity. The superposition of a group of such waves to form a localized packet occurs by a subtle phase interference that is, on the whole, constructive only over a localized region in space and destructive throughout the remainder of space. The phase interference changes with time if the component waves move at different speeds, so the spatial peak in the localized shape can move over time, even relative to a given point on the moving component wave that has a k-value exactly equal to that of the center of the band of superimposed waves.

If a packet is moving in some direction in space that is not parallel to the x-axis of our coordinate system, then the group velocity will have components along the three axes of the coordinate system. The dispersion relation will take the form $\omega = \omega(\mathbf{k})$, where $\mathbf{k}$ is a vector pointing in a given direction. Since ω is a scalar quantity, $\omega(\mathbf{k})$ maps out a type of three-dimensional surface for the dependence of frequency on direction and magnitude of the $\mathbf{k}$-vector. The $\mathbf{k}$-vector represents the direction of propagation of one component wave of the packet, and the magnitude of the $\mathbf{k}$-vector still has the interpretation of being $2\pi/\lambda$, where λ is the wavelength of the component wave in its propagation direction. The group velocity components in the x, y, and z directions in space can be obtained by generalizing Eq. (96), which was derived on the assumption that there was only x-motion. A proper way to carry out the derivation is to use the Taylor series expansion of $\omega(\mathbf{k})$ in three dimensions. Three first-derivative terms can be obtained in place of one, and the multipliers of k_x, k_y, and k_z are, respectively, $[x - v_{\text{group}}^{(x)}t]$, $[y - v_{\text{group}}^{(y)}t]$, and $[z - v_{\text{group}}^{(z)}t]$, where

$$v_{\text{group}}^{(x)} = \left(\frac{\partial\omega(\mathbf{k})}{\partial k_x}\right)_{\mathbf{k}=\mathbf{k}_0} \tag{104}$$

$$v_{\text{group}}^{(y)} = \left(\frac{\partial\omega(\mathbf{k})}{\partial k_y}\right)_{\mathbf{k}=\mathbf{k}_0} \tag{105}$$

$$v_{\text{group}}^{(z)} = \left(\frac{\partial\omega(\mathbf{k})}{\partial k_z}\right)_{\mathbf{k}=\mathbf{k}_0}. \tag{106}$$

The conclusion to be reached is that the group velocity is a vector with components given by Eqs. (104)–(106). These results can be abbreviated by writing

$$\mathbf{v}_{\text{group}} = \nabla_{\mathbf{k}}\omega(\mathbf{k})|_{\mathbf{k}=\mathbf{k}_0} \tag{107}$$

as the general expression for the group velocity for wave motion in three-dimensional space.

V. WAVE–PARTICLE DUALITY

A. Ideas of de Broglie

The discovery that particles diffract like waves was one of the most important in physics, since the entire discipline of quantum mechanics is based on a wave description of matter. Preceding that discovery, de Broglie carried out an analysis of the hypothetical wave properties of matter by employing special relativity theory, with an insightful hunch that matter is not all that different in its fundamental wave behavior from electromagnetic radiation. His idea was that since radiation had been shown to have both wavelike and particlelike properties, perhaps nonzero rest mass "particles" also could manifest both wavelike and particlelike properties. However, different experimental conditions might be required for matter to manifest one property or the other. De Broglie called his idea wave–particle duality.

In 1926 Walter Maurice Elsasser suggested that the de Broglie hypothesis could be tested by aiming an electron beam at the surface of a crystalline solid to see if diffraction spots were obtained in the same way as observed in X-ray diffraction. Experiments were carried out shortly thereafter by Davisson and Germer and also by G. P. Thomson, so that by 1927, the de Broglie hypothesis had been experimentally verified. Thus, the electron discovered by J. J. Thomson in 1987 enjoyed its status as a classical particle for a period of less than 30 years before its schizophrenic nature surfaced. Amazingly, it was later confirmed not only that electrons had this dual nature but also that other particles, including neutrons and atoms, could be diffracted. Particles somehow have the inherent property of being able to experience mutual interference even while maintaining such an extremely high degree of localization in space that they would not appear to have any physical overlap. What boggles the mind even more, single particles themselves behave statistically in accordance with the same diffraction probability distribution.

Because electromagnetic radiation can manifest discreteness properties, with the photon energy $\mathscr{E}$ related to frequency according to

$$\mathscr{E} = h\nu = \hbar\omega \tag{108}$$

de Broglie reasoned that, in analogy, any wave properties of particles would also have some associated frequency. Moreover, in accordance with the concept of wave–particle duality and the similarity of particle and electromagnetic radiation in nature, he assumed that the energy–frequency relation for particles would be the same as the energy–frequency relation for photons given by Eq. (108).

The term particle generally is utilized to refer to those fundamental entities for which the rest mass m_0 is nonzero. Exactly what the frequency of a particle refers to is not known, but it can be imagined that there is some internal mode of oscillation.

B. Development of the de Broglie Relation

If a particle can ever be localized to any extent in space, as there is certainly every right to expect, and simultaneously is to have its experimentaly observed wavelike character described in some fashion by waves, then it is quite logical to consider a wave packet of the sort treated in Section IV.H. The waves forming the packet necessarily would be matter waves, but the diffraction results lead to the belief that these waves have the same linear superposition properties of all other familiar types of waves. The waves must be associated with the presence of the particle in some way, so it is reasonable to expect that the group velocity $\mathbf{v}_{\text{group}}$ of the wave packet will be equal to the particle velocity $\mathbf{v}_{\text{particle}}$, or

$$\mathbf{v}_{\text{group}} = \mathbf{v}_{\text{particle}} = \frac{\mathbf{p}}{m}, \tag{109}$$

where $\mathbf{p}$ is the momentum of the particle of mass m. Let us for the moment confine our attention to a single direction in space. The discussion will be generalized to three dimensions later, whenever there are important implications.

The group velocity expression $v_{\text{group}} = dw/dk$ given by Eq. (96), which was derived in Section IV.H, together with the derivative of the energy–frequency relation given by Eq. (108), postulated by de Broglie, can be utilized to obtain

$$d\mathscr{E} = \hbar\, d\omega = \hbar\, v_{\text{group}}\, dk. \tag{110}$$

Next, the energy–momentum relation

$$\mathscr{E} = \left(\mathscr{E}_0^2 + p^2c^2\right)^{1/2} \tag{111}$$

given by Eq. (22) is differentiated and the mass–energy relation given by Eq. (20) in Section II on the special theory of relativity is used to obtain a second independent expression for $d\mathscr{E}$,

$$\begin{aligned} d\mathscr{E} &= \frac{c^2p\, dp}{\left(\mathscr{E}_0^2 + p^2c^2\right)^{1/2}} = \frac{c^2p\, dp}{mc^2} \\ &= \frac{p}{m}\, dp = v_{\text{particle}}\, dp. \end{aligned} \tag{112}$$

Equating the two independent expressions given for $d\mathscr{E}$ by Eqs. (110) and (112) gives

$$v_{\text{particle}}\, dp = v_{\text{group}}\hbar\, dk. \tag{113}$$

Identifying the group velocity with the particle velocity, according to Eq. (109), and dividing Eq. (113) by this quantity gives

$$dp = \hbar\, dk. \tag{114}$$

This remarkable result relates the change in the reciprocal of the wavelength of a particle matter–wave component to a change in the momentum of the particle. Since by definition $k = 2\pi/\lambda$, the relation given by Eq. (114) also can be written in the physically more meaningful form,

$$dp = \frac{h}{2\pi}\, d\left(\frac{2\pi}{\lambda}\right) = -\frac{h}{\lambda^2}\, d\lambda. \tag{115}$$

Integrating Eqs. (114) and (115) from any arbitrary reference point, denoted by subscript 0, gives

$$p - p_0 = \hbar k - \hbar k_0 = h\left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right). \tag{116}$$

Thus

$$p = \hbar k = \frac{h}{\lambda}, \tag{117}$$

which derives the de Broglie relation, a relationship between the momentum of a particle and the wavelength of the matter wave associated with a particle having this precisely defined momentum. Built into this development is the fundamental hypothesis of de Broglie that matter has associated with it a frequency that is related to the total energy in the same way that the frequency of electromagnetic radiation is related to the individual photon energy.

If the photon is considered to be a quantum particle in every respect, except for being a limiting case from the standpoint of having zero rest mass, then

$$p_{\text{photon}} = \hbar k_{\text{photon}} = \frac{h}{\lambda_{\text{photon}}} \tag{118}$$

As an alternative, applying the energy–momentum relation given by Eq. (22) to the case of photons, which have zero rest mass and, hence, zero rest-mass energy in accordance with $\mathscr{E}_0 = m_0c^2$, gives

$$\mathscr{E}_{\text{photon}} = \pm p_{\text{photon}}c. \tag{119}$$

The sign of the root is chosen to give a positive value for the photon energy, since there is no physical interpretation for negative photon energy at the present moment.

Next, let us use the energy–frequency relation given by Eq. (108) to obtain an independent relation for the photon energy:

$$\begin{aligned} \mathscr{E}_{\text{photon}} &= \hbar\omega_{\text{photon}} = \hbar(2\pi\nu_{\text{photon}}) \\ &= h\nu_{\text{photon}} = \frac{hc}{\lambda_{\text{photon}}}. \end{aligned} \tag{120}$$

Equating the two independent results of Eqs. (119) and (120) for the photon energy and dividing through by the light velocity c gives

$$p_{\text{photon}} = \frac{h}{\lambda_{\text{photon}}}. \tag{121}$$

This relation is consistent with Eq. (118) and the experimentally proven hypothesis that light has a momentum associated with it. This fact is also a part of classical electromagnetic theory, but the form is somewhat different from the quantum result given here.

As is evident from the development of Eq. (117), the relation is valid for particles of any rest mass, including the limiting case of zero rest mass, and it is valid at any velocity, including the limits of low velocity and relativistic velocity. This relation has been confirmed experimentally by accelerating electrons to different velocities, including relativistic velocities.

It seems a bit strange that such a fundamental relation of quantum mechanics is deduced from the theory of special relativity, considering the discussion in Section II.B of the difference in philosophical bases for the two theories. To be specific, the fundamental equations of a deterministic theory have been used here to deduce a wavelength relationship for particles, the preciseness of this wavelength determining to a large extent the inherent uncertainty in the specification of the location of the particle. One should not be misled, however, into believing that relativity theory has *predicted* the wave nature of particle. Instead, the wave nature of particles is the initial hypothesis that, in itself, is supported by experimental observation that particles diffract from crystal lattices in a wavelike manner.

C. Phase Velocity for Free Particles

The relations $\omega = \mathcal{E}/\hbar$ and $k = p/\hbar$ can be used to evaluate the phase velocity v_{phase} of matter waves:

$$v_{\text{phase}} = \frac{\omega}{k} = \frac{\hbar\omega}{\hbar k} = \frac{\mathcal{E}}{p} = \frac{mc^2}{mv_{\text{particle}}} = \frac{c^2}{v_{\text{particle}}}. \qquad (122)$$

The phase velocity of matter waves thus depends upon the particle velocity. As the particle speed increases, the phase velocity decreases.

Although the phase velocity takes its most natural form when expressed in term of the particle velocity, it also can be expressed in terms of the particle momentum:

$$v_{\text{phase}} = \frac{\omega}{k} = \frac{\mathcal{E}}{p} = p^{-1}\left(\mathcal{E}_0^2 + p^2c^2\right)^{1/2} = c\left[1 + \left(\frac{p_{\text{C}}}{p}\right)^2\right]^{1/2}, \qquad (123)$$

where $p_{\text{C}} = m_0c$. This also can be expressed as

$$v_{\text{phase}} = c\left[1 + \left(\frac{k_{\text{C}}}{k}\right)^2\right]^{1/2} = c\left[1 + \left(\frac{\lambda}{\lambda_{\text{C}}}\right)^2\right]^{1/2}, \qquad (124)$$

where $k_{\text{C}} = p_{\text{C}}/\hbar = m_0c/\hbar$ and $\lambda_{\text{C}} = 2\pi/k_{\text{C}} = h/m_0c$. The parameter λ_{C} is called the Compton wavelength whenever m_0 is the electron rest mass. It is not the wavelength of an electron at rest, of course, since the de Broglie relation yields infinity as the wavelength of a particle with zero momentum.

For the special case of photons and other particles that have zero rest mass, the phase velocity is given by

$$v_{\text{phase}}\Big|_{m_0=0} = \frac{\mathcal{E}}{p} = \frac{1}{p}\left[m_0c^2 + p^2c^2\right]^{1/2}\Big|_{m_0=0} = \frac{pc}{p} = c. \qquad (125)$$

Matter waves have phase velocities ranging upward from c, with the phase velocity approaching infinity as the particle speed approaches zero. The phase velocity represents the speed of motion of a single, isolated and completely extended component wave having no position modulation of its shape and, by itself, can carry no information. Thus, there is no conflict with the tenet of special relativity that a signal cannot travel with a velocity exceeding the speed of light.

To summarize for matter waves, the group velocity must be equal to the particle velocity and the phase velocity varies inversely with the particle velocity. The critical quantity for evaluating the wavelength is the momentum, whereas the critical quantity for evaluating the phase velocity is the particle velocity. Particles having different rest-mass values have the same wavelength whenever the velocities differ such as to give the same values for the momentum. Particles having different rest-mass values have the same phase velocity, however, whenever they have the same value for the particle velocity.

D. Dispersion Relation for Free Particles

It has been shown in Section IV that the dispersion relation for light is $\omega = ck$. This immediately gives the phase velocity $\omega/k = c$ and the group velocity $d\omega/dk = c$.

The dispersion relation for free particles is deduced by utilizing the energy–momentum relation from special relativity and substituting $\mathcal{E} = \hbar\omega$ and $p = \hbar k$, so that

$$\hbar\omega = \left(\mathcal{E}_0^2 + \hbar^2k^2c^2\right)^{1/2}. \qquad (126)$$

This free-particle dispersion relation also can be written in the form

$$\omega = \omega_0\left[1 + \left(\frac{\hbar}{m_0c}\right)^2 k^2\right]^{1/2}, \qquad (127)$$

where $\omega_0 = m_0c^2/\hbar$, the frequency corresponding to the rest mass m_0. In terms of parameters $\lambda_{\text{C}} = h/m_0c$ and $k_{\text{C}} = 2\pi/\lambda_{\text{C}}$, Eq. (127) becomes

$$\omega = ck_{\text{C}}\left[1 + \left(\frac{k}{k_{\text{C}}}\right)^2\right]^{1/2}. \qquad (128)$$

Binomial expansion illustrates a quadratic dependence of ω on k with a cutoff in frequency below the rest-mass frequency ω_0:

$$\omega = \omega_0\left[1 + \frac{k^2}{2k_{\text{C}}^2} + \cdots\right]. \qquad (129)$$

For low values of the particle momentum, k is small and the frequency does not depart very much from the rest-mass frequency. This is merely a reflection of the fact that $\mathscr{E} = mc^2 = \gamma m_0 c^2 = \hbar\omega$ predicts that

$$\frac{\omega}{\omega_0} = \gamma, \tag{130}$$

where γ is the parameter defined by Eq. (19) in Section II on special relativity. This might be described by considering the particle as initially created in a zero potential energy region with a proper frequency, which it then maintains always during its lifetime. Any perceived difference in frequency due to motion is then merely due to the relativistic shift in energy with Galilean reference frame. This is analogous to the increase in mass with the particle speed; that is, $m = \gamma m_0$, so $\mathscr{E} = mc^2 = \gamma m_0 c^2$. Thus, $\mathscr{E} = \hbar\omega = \gamma\hbar\omega_0$, thereby giving $\omega = \gamma\omega_0$ in Eq. (130). This does maintain the de Broglie premise that a particle has a frequency associated with it even when it is stationary, that frequency being dependent upon the rest mass m_0 in accordance with the relation $\mathscr{E}_0 = m_0 c^2 = \hbar\omega_0$.

E. Energy–Momentum Relations for Particles with Potential Energy

It may be appreciated that developing the dispersion relation $\omega(k)$ versus k can be sufficient preparation for evaluating the phase velocity $v_{\text{phase}} = \omega(k)/k$, according to Eq. (69). Because most applications of quantum mechanics involve particles with some type of potential energy, it is necessary to consider this situation.

A force $\mathbf{F}$ changes the momentum $\mathbf{p}$ of a free particle in accordance with $\mathbf{F} = d\mathbf{p}/dt$, or, equivalently, the change $d\mathbf{p}$ in particle momentum is given by $\mathbf{F}\,dt$. Since the work done by a force in moving a particle through a vector distance $d\mathbf{r}$ is $dW = \mathbf{F}\cdot d\mathbf{r}$, the power input $\mathscr{P} = dW/dt$ from the source is $\mathbf{F}\cdot d\mathbf{r}/dt = \mathbf{F}\cdot\mathbf{v}$. In one dimension, $\mathscr{P} = Fv$. In our case, the power to change the motion energy of the particle comes from the internal potential energy. For a conservative system, the potential energy change with position leads to an internal force on the particle

$$\mathbf{F} = -\nabla U(\mathbf{r}) \tag{131}$$

which, for one dimension, is simply

$$F = \frac{-dU(r)}{dr}. \tag{132}$$

Thus

$$\frac{dp}{dt} = F = \frac{-dU(r)}{dr} \tag{133}$$

or, equivalently

$$dU(r) = \frac{-dp}{dt}\,dr. \tag{134}$$

But $v_{\text{particle}} = dr/dt$, so $dr = v_{\text{particle}}\,dt$, and the result is

$$\begin{aligned} dU(r) &= \frac{-dp}{dt}\mathrm{v}_{\text{particle}}\,dt = -v_{\text{particle}}\,dp \\ &= \frac{-p}{m}\,dp \end{aligned} \tag{135}$$

This relation cannot be integrated directly because m depends upon particle velocity and, hence, upon p. However, substituting $m = \mathscr{E}/c^2$, with $\mathscr{E}$ given by the usual energy–momentum relation of special relativity in Eq. (22), leads to a perfect differential:

$$\begin{aligned} dU(r) &= -(c^2 p\,dp)\left(\mathscr{E}_0^2 + p^2c^2\right)^{-1/2} \\ &= -d\left(\mathscr{E}_0^2 + p^2c^2\right)^{1/2}. \end{aligned} \tag{136}$$

Integrating from a classical turining point, defined as a position where $p = 0$ so the potential energy at the point is the total energy $\mathscr{E}_{\mathrm{T}}^{(\mathscr{N}t)}$ in classical Newtonian physics, we obtain

$$U(r) - \mathscr{E}_{\mathrm{T}}^{(\mathscr{N}t)} = m_0c^2 - \left(\mathscr{E}_0^2 + p^2c^2\right)^{1/2}. \tag{137}$$

The classical total energy is conserved as long as no rest mass is created or annihilated. A new constant $\mathscr{E}_{\mathrm{T}}^{(\mathscr{R}e\ell)}$ thus can be defined as

$$\mathscr{E}_{\mathrm{T}}^{(\mathscr{R}e\ell)} = \mathscr{E}_{\mathrm{T}}^{(\mathscr{N}t)} + m_0c^2 \tag{138}$$

and Eq. (137) can be written as

$$\mathscr{E}_{\mathrm{T}}^{(\mathscr{R}e\ell)} - U(r) = \left(\mathscr{E}_0^2 + p^2c^2\right)^{1/2}. \tag{139}$$

This is the energy–momentum relation for a particle having a potential energy. Note from this relation that when the potential energy is zero, $\mathscr{E}_{\mathrm{T}}^{(\mathscr{R}e\ell)}$ becomes the total relativistic energy $\mathscr{E}$ of a free particle. The constant $\mathscr{E}_{\mathrm{T}}^{(\mathscr{R}e\ell)}$, in lieu of $\mathscr{E}$, is the conserved quantity when a potential energy is included. Let us also define a quantity $\mathscr{E}_{\mathrm{K}}^{(\mathscr{R}e\ell)}$

$$\mathscr{E}_{\mathrm{K}}^{(\mathscr{R}e\ell)} = \mathscr{E}_{\mathrm{T}}^{(\mathscr{N}t)} - U(r) \tag{140}$$

or, alternatively, $\mathscr{E}_{\mathrm{T}}^{(\mathscr{N}t)} = \mathscr{E}_{\mathrm{K}}^{(\mathscr{R}e\ell)} + U(r)$. By employing Eq. (138), Eq. (140) can be written in the form

$$\begin{aligned} \mathscr{E}_{\mathrm{K}}^{(\mathscr{R}e\ell)} &= \left[\mathscr{E}_{\mathrm{T}}^{(\mathscr{R}e\ell)} - m_0c^2\right] - U(r) \\ &= \left[\mathscr{E}_{\mathrm{T}}^{(\mathscr{R}e\ell)} - U(r)\right] - m_0c^2 \end{aligned} \tag{141}$$

and, by employing Eq. (139), this expression can then be written in the form

$$\mathscr{E}_{\mathrm{K}}^{(\mathscr{R}e\ell)} = \left(\mathscr{E}_0^2 + p^2c^2\right)^{1/2} - m_0c^2. \tag{142}$$

Equation (141), in the form

$$\mathscr{E}_{\mathrm{T}}^{(\mathscr{R}e\ell)} = \mathscr{E}_{\mathrm{K}}^{(\mathscr{R}e\ell)} + U(r) + m_0c^2 \tag{143}$$

seems intuitive from a scalar-energy viewpoint. This form has its primary usefulness in the low-velocity limit. In the relativistic limit, the best form for $\mathscr{E}_{\mathrm{T}}^{(\mathscr{R}e\ell)}$ is that obtained from Eq. (139):

$$\mathscr{E}_{\mathrm{T}}^{(\mathscr{R}e\ell)} = \left(\mathscr{E}_0^2 + p^2c^2\right)^{1/2} + U(r) \qquad (144)$$

since this form is intuitive from a free-particle viewpoint.

Squaring Eq. (139) and solving for p^2 gives

$$p^2 = \frac{1}{c^2}\left\{\left[\mathscr{E}_{\mathrm{T}}^{(\mathscr{R}e\ell)} - U(r)\right]^2 - \mathscr{E}_0^2\right\}. \qquad (145)$$

Substituting Eq. (141) then gives

$$p^2 = \frac{1}{c^2}\left\{\left[\mathscr{E}_{\mathrm{K}}^{(\mathscr{R}e\ell)} + m_0c^2\right]^2 - \left(m_0c^2\right)^2\right\}, \qquad (146)$$

which relates the momentum to the quantity $\mathscr{E}_{\mathrm{K}}^{(\mathscr{R}e\ell)}$. Rearranging Eq. (146) to obtain $\mathscr{E}_{\mathrm{K}}^{(\mathscr{R}e\ell)}$ explicitly in terms of p gives

$$\mathscr{E}_{\mathrm{K}}^{(\mathscr{R}e\ell)} = -m_0c^2 + m_0c^2\left|\left(1 + \frac{p^2c^2}{m_0^2c^4}\right)^{1/2}\right|. \qquad (147)$$

Employing the binomial expansion then gives the series approximation

$$\mathscr{E}_{\mathrm{K}}^{(\mathscr{R}e\ell)} \simeq \frac{p^2}{2m_0}\left(1 - \frac{p^2}{4m_0^2c^2} + \cdots\right). \qquad (148)$$

In the low-velocity limit defined by $\gamma^2 \simeq 1$, which requires $(v/c)^2 \ll 1$ [see Eq. (19)], $p \simeq m_0 v$, and the quantity $\mathscr{E}_{\mathrm{K}}^{(\mathscr{R}e\ell)}$ thus reduces in lowest order to the Newtonian expression for the kinetic energy

$$\mathscr{E}_{\mathrm{K}}^{(\mathscr{N}t)} = \frac{1}{2}m_0v^2. \qquad (149)$$

F. Dispersion Relations for Particles with Potential Energy

To develop a wave dispersion relation applicable to particles having a potential energy, the wavelength is again associated with the particle momentum by means of the de Broglie relation $\lambda = h/p$. Equivalently, $p = h/\lambda = (h/2\pi)(2\pi/\lambda) = \hbar k$, in accordance with Eq. (117).

According to the de Broglie hypothesis, a frequency ω must be associated with the particle energy. Although the nature of the frequency of a particle is not yet understood, a guideline can be the familiar results from the particular limiting case of a zero rest-mass quantum particle—namely, the photon. One must remain alert, however, to avoid introducing subtle errors in using the analogy.

The photon has a wavelength λ and a frequency ν. As the photon travels in free space, it has a speed $c = \lambda\nu$. When the photon enters a dielectric material medium of refractive index n, its speed decreases to c/n but its frequency ν does not change. The photon energy $\mathscr{E} = h\nu$ thus does not change. The wavelength decreases in accordance with $\lambda_{\mathrm{pho}}^{\mathrm{med}} = \lambda_{\mathrm{pho}}^{\mathrm{vac}}/n$, where $\lambda_{\mathrm{pho}}^{\mathrm{med}}$ is the wavelength in the medium and $\lambda_{\mathrm{pho}}^{\mathrm{vac}}$ is the wavelength in free space. The photon velocity is still given by the product of photon frequency and the appropriate photon wavelenght, whether the medium be a dielectric or free space. The momentum $p_{\mathrm{pho}}^{\mathrm{vac}} = h/\lambda_{\mathrm{pho}}^{\mathrm{vac}}$ of the photon in free space changes to $p_{\mathrm{pho}}^{\mathrm{med}} = h/\lambda_{\mathrm{pho}}^{\mathrm{med}}$ as it enters the dielectric medium. Thus, the constant of motion for the photon is the frequency $\nu = \omega/2\pi$, with wavelength λ and momentum p being medium-dependent. It is an interesting observation that, from the viewpoint of Eq. (142), all energy of the photon is kinetic.

One might be tempted to assume, in analogy with the photon, that the frequency of a particle located in a purely conservative potential energy region is a constant of motion. This would give rise to the picture of a particle entering a region of varying potential energy, where it could be accelerated. Its speed would change with its acceleration. Its momentum would change, so its wavelength would change also. However, the *total* energy of the particle would not change. Both remaining unchanged, the frequency would then maintain a direct relationship to the total energy. The analogy with the behavior of a photon as it enters a region of varying refractive index would, in this way, be exploited to the fullest. However, this might seem to do great violence to the earlier conclusion that the frequency of the wave associated with a free particle depends upon the speed of the particle. In the present case, one would be considering the particle changing speed due to the change in internal potential energy as it moves through the potential energy region, yet one would be proposing to hold its frequency ω to be a constant. Moreover, the constant frequency would not be equal to the *proper* frequency ω_0 for the particle in the relativistic sense but would have some value that would necessarily be related to $U(r)$, since even the choice of the zero point for measuring $U(r)$ would affect the value of ω.

This strange state of affairs might be viewed as unacceptable. The seeming enigma can be resolved, however, in the following way. The total energy of a particle in a potential energy $U(\mathbf{r})$ can be viewed, apart from some constant reference energy, as the energy of a quantum system consisting of the particle in question plus the source medium responsible for the potential energy. As the particle moves about in the potential-energy region, its kinetic and potential energy changes in such a way that the total energy remains invariant. Thus, an increase in the kinetic energy of the particle comes at the expense of its potential energy, so the chemical binding energy of the system is increased. Similarly, a decrease in the kinetic energy of the particle means that its potential energy is algebraically increased, which corresponds to a decrease in the chemical binding energy of the system. The relationship between energy and frequency for the particle in the system can be postulated to involve the kinetic energy of the particle and the attendant local modification in the binding energy of the system. This is consistent with choosing $\hbar\omega = \mathscr{E}_{\mathrm{T}}^{(\mathscr{R}e\ell)}$ instead of $\mathscr{E}_{\mathrm{T}}^{(\mathscr{R}e\ell)} - U(r)$, so that

$$\omega = \frac{\mathscr{E}_{\mathrm{T}}^{(\mathscr{R}e\ell)}}{\hbar}, \tag{150}$$

where $\mathscr{E}_{\mathrm{T}}^{(\mathscr{R}e\ell)}$ is defined by Eq. (144).

G. Phase Velocity of Particles with Potential Energy

Taking the frequency of the particle wave from Eq. (150) and using the de Broglie relation $k = p/\hbar$ given by Eq. (117) leads to an evaluation of the phase velocity

$$v_{\text{phase}} = \frac{\omega}{k} = \frac{\hbar\omega}{\hbar k} = \frac{\mathscr{E}_{\mathrm{T}}^{(\mathscr{R}e\ell)}}{p}. \tag{151}$$

VI. WAVE EQUATIONS FOR PARTICLES

A. Master Equations for Particles with Potential Energy

In this section, it is assumed at the outset that the reader already has studied carefully the material presented earlier, especially the fundamentals of classical wave motion in Section IV and the ideas of wave–particle duality in Section V. This section relies heavily on the concepts of dispersion relations and phase velocity in wave motion covered in those sections.

Utilizing the three-dimensional classical wave equation [Eq. (57)] and substituting Eq. (151) for the phase velocity gives the wave equation for particles

$$\nabla^2\Psi = \left[\frac{p}{\mathscr{E}_{\mathrm{T}}^{(\mathscr{R}e\ell)}}\right]^2 \left(\frac{\partial^2\Psi}{\partial t^2}\right) = \mathscr{R}\frac{\partial^2\Psi}{\partial t^2} \tag{152}$$

or, equivalently

$$\frac{1}{\mathscr{R}}\nabla^2\Psi = \frac{\partial^2\Psi}{\partial t^2} \tag{153}$$

with $\mathscr{R} \equiv [p/\mathscr{E}_{\mathrm{T}}^{(\mathscr{R}e\ell)}]^2$. It should be noted from Eq. (145) for p that $\mathscr{R}$ will depend upon position $\mathbf{r}$, but it does not depend explicitly upon the time t. Therefore, the variables can be separated in Eq. (152), so a solution can be carried out by the separation-of-variables technique. Substituting the product trial solution

$$\Psi(\mathbf{r}, t) = \Phi(\mathbf{r})T(t) \tag{154}$$

into Eq. (153) and dividing through by ψ yields

$$\left(\frac{1}{\mathscr{R}}\right)\left(\frac{1}{\Phi}\right)\nabla^2\Phi = \left(\frac{1}{T}\right)\left(\frac{\partial^2 T}{\partial t^2}\right). \tag{155}$$

The usual argument for the separation of variables holds. The right-hand side is a function (at most) of the variable t, the left-hand side is a function (at most) of the variable $\mathbf{r}$. Since the two sides are required to be equal, then neither can vary with t or $\mathbf{r}$. Setting the right-hand side equal to the constant Γ and employing the trial solution

$$T(t) = T_0 \exp(-i\Omega t) \tag{156}$$

yields a solution, provided that

$$\Gamma = -\Omega^2. \tag{157}$$

Since T_0 is unspecified, it can be conveniently chosen to be unity. The right-hand side of Eq. (155) equals Γ, so the left-hand side must equal Γ also. Thus, the differential equation for $\Phi(\mathbf{r})$ is

$$\left(\frac{1}{\mathscr{R}}\right)\left(\frac{1}{\Phi}\right)\nabla^2\Phi = -\Omega^2 \tag{158}$$

or, equivalently

$$\nabla^2\Phi(\mathbf{r}) + \Omega^2\mathscr{R}\Phi(\mathbf{r}) = 0. \tag{159}$$

There still remains the task of evaluating the constant Ω. The approach used is to consider the limiting case of a constant potential energy $U(\mathbf{r}) = U_0$, so that the momentum $\mathbf{p}$ is a constant. There is then no position-dependence of $\mathscr{R}$. The quantity Ω^2 is always a constant [cf. Eq. (157)], so that the product $\Omega^2\mathscr{R}$ is then a constant. As a consequence, Eq. (159) has particularly simple solutions. The trial solution

$$\begin{aligned}\Phi &= \Phi_0 \exp(i\mathbf{k}\cdot\mathbf{r})\\ &= \Phi_0 \exp[i(k_x x + k_y y + k_z z)]\end{aligned} \tag{160}$$

satisfies the equation, with $\mathbf{k}$ being a vector constant. This solution represents a wave having wavelength $2\pi/k$, with $k = (k_x^2 + k_y^2 + k_z^2)^{1/2}$, which is traveling in the direction of the fixed vector $\mathbf{k}$. In one dimension (e.g., a particle moving along the x-direction), the exponential function reduces simply to $\exp(ikx)$. The k-vector involves the wavelength for the particle wave and therefore is associated with the particle momentum p in accordance with the de Broglie relation $\mathbf{p} = \hbar\mathbf{k}$. The momentum in this case of a constant potential energy is a constant of motion. Substituting this solution into Eq. (159) gives

$$-k^2\Phi = \Omega^2\mathscr{R}\Phi = 0 \tag{161}$$

so that

$$\Omega^2 = \frac{k^2}{\mathscr{R}} = \left(\frac{k^2}{p^2}\right)[\mathscr{E}_{\mathrm{T}}^{(\mathscr{R}e\ell)}]^2. \tag{162}$$

Since $\mathbf{p} = \hbar\mathbf{k}$, this leads to the conclusion that

$$\Omega = \frac{\mathscr{E}_{\mathrm{T}}^{(\mathscr{R}e\ell)}}{\hbar}. \tag{163}$$

It can be noted from comparison of this result with Eq. (150) that the separation constant $\Gamma = -\Omega^2$ is equal to $-\omega^2$. Since, for a conservative system, $\mathscr{E}_{\mathrm{T}}^{(\mathscr{R}e\ell)}$ is a constant, independent of whether or not $U(\mathbf{r})$ is a function of $\mathbf{r}$, it may be concluded that Eq. (163) provides a reasonable result for Ω for any arbitrary form of $U(\mathbf{r})$. This form converts Eq. (159) to the form

$$\nabla^2\Phi(\mathbf{r}) + \left[\frac{\mathscr{E}_{\mathrm{T}}^{(\mathscr{R}e\ell)}}{\hbar}\right]^2 \mathscr{R}\Phi(\mathbf{r}) = 0 \tag{164}$$

or, equivalently, by using the definition of $\mathscr{R}$

$$\nabla^2\Phi(\mathbf{r}) + \frac{p^2}{\hbar^2}\Phi(\mathbf{r}) = 0. \tag{165}$$

An alternate form of this equation involving the potential energy $U(\mathbf{r})$ is obtained by utilizing Eq. (145) for p^2, which leads to what is referred to herein as the "master equation"

$$\nabla^2\Phi(\mathbf{r}) + \left(\frac{1}{\hbar c}\right)^2 \left\{\left[\mathscr{E}_{\mathrm{T}}^{(\mathscr{R}e\ell)} - U(\mathbf{r})\right]^2 - \mathscr{E}_0^2\right\}\Phi(\mathbf{r}) = 0. \tag{166}$$

B. Approximation to the Master Equation in the Nonrelativistic Limit

In the nonrelativistic limit

$$p^2 = 2m_0\left[\mathscr{E}_{\mathrm{T}}^{(\mathscr{N}t)} - U(\mathbf{r})\right], \tag{167}$$

where $\mathscr{E}_{\mathrm{T}}^{(\mathscr{N}t)}$ is the Newtonian total energy discussed in Section V.E. Thus, Eq. (165) reduces in the nonrelativistic limit to the form

$$\nabla^2\phi(\mathbf{r}) + \left(\frac{2m_0}{\hbar^2}\right)\left[\mathscr{E}_{\mathrm{T}}^{(\mathscr{N}t)} - U(\mathbf{r})\right]\phi(\mathbf{r}) = 0, \tag{168}$$

where ϕ is used instead of Φ to indicate the nonrelativistic result. This represents the general nonrelativistic equation for the motion of a particle in a region of variable potential energy.

C. The Schrödinger Equation

Equation (168), known as the time-independent Schrödinger equation, is usually written in the form

$$-\frac{\hbar^2}{2m_0}\nabla^2\phi_j + U(\mathbf{r})\phi_j = \mathscr{E}_j\phi_j, \tag{169}$$

where the subscript denotes a set of possible solutions ϕ_j with attendant discrete values for the Newtonian total energy $\mathscr{E}_{\mathrm{T}}^{(\mathscr{N}t)}$. The ϕ_j solutions are called the Schrödinger time-independent energy eigenfunctions. Equation (169) is an energy eigenvalue equation, since it can be written in the form

$$\mathscr{E}_{\mathrm{T}}^{(op)}\phi_j(\mathbf{r}) = \mathscr{E}_j\phi_j(\mathbf{r}) \tag{170}$$

with the total energy operator $\mathscr{E}_{\mathrm{T}}^{(op)}$ defined by

$$\mathscr{E}_{\mathrm{T}}^{(op)} = -\frac{\hbar^2}{2m_0}\nabla^2 + U(\mathbf{r}). \tag{171}$$

Usually, $\mathscr{E}_{\mathrm{T}}^{(op)}$ is denoted by $\mathscr{H}$ and is called the Hamiltonian operator. It is of the nature of an eigenvalue equation that an operator representing some physical quantity operates on a functions known as the eigenfunction [in this case, $\phi_j(\mathbf{r})$] to produce the product of a constant with the eigenfunction. The constant is the eigenvalue. It represents the constant of motion associated with the physical state represented by the eigenfunction.

Let us define an oscillatory, time-dependent function involving the Newtonian total energy $\mathscr{E}_{\mathrm{T}}^{(\mathscr{N}t)} = \mathscr{E}_j = \hbar\omega_j$ for the particle

$$\theta_j(t) = \theta_0 \exp\left(-\frac{i}{\hbar}\mathscr{E}_j t\right) = \theta_0\exp(-i\omega_j t). \tag{172}$$

Multiplying Eq. (169) through by this factor leads to the equation

$$-\frac{\hbar^2}{2m_0}\nabla^2\psi_j + U(\mathbf{r})\psi_j = \mathscr{E}_j\psi_j, \tag{173}$$

where

$$\psi_j = \phi_j(\mathbf{r})\theta_j(t). \tag{174}$$

It can be noted from Eq. (154) and the subsequent development that with $T_0 = 1$, for a selected energy solution denoted by the subscript i, $\Psi_i = \Phi_i(\mathbf{r})\exp[-i/\hbar)\mathscr{E}_{\mathrm{T}}^{(\mathscr{R}e\ell)}t]$. In Eq. (174), with the choice $\theta_0 = 1$, the corresponding nonrelativistic solution is given by $\psi_i = \phi_i(\mathbf{r})\exp[-(i/\hbar)\mathscr{E}_{\mathrm{T}}^{(\mathscr{N}t)}t]$. The ratio of the two wavefunctions is

$$\frac{\Psi_i}{\psi_i} = \frac{\Phi_i}{\phi_i}\exp\left(-\frac{i}{\hbar}\mathscr{E}_0 t\right), \tag{175}$$

where $\mathscr{E}_0$ is the rest mass energy m_0c^2. The ψ_j are called the Schrödinger time-dependent eigenfunctions. Note that

$$i\hbar\frac{\partial}{\partial t}\psi_j = \mathscr{E}_j\psi_j \tag{176}$$

so that Eq. (173) can be written in the form

$$-\frac{\hbar^2}{2m_0}\nabla^2\psi_j + U(\mathbf{r})\psi_j = i\hbar\frac{\partial\psi_j}{\partial t}. \tag{177}$$

This is a linear equation, so a general solution ψ can be constructed by superposition of all individual linearly independent solutions

$$\psi = \sum_j a_j\psi_j, \tag{178}$$

where the a_j are arbitrary complex constants. This is called the Schrödinger wavefunction. The equation analogous to Eq. (177) representing the most general solution is thus

$$-\frac{\hbar^2}{2m_0}\nabla^2\psi + U(\mathbf{r})\psi = i\hbar\frac{\partial\psi}{\partial t}. \qquad (179)$$

This is called the time-dependent Schrödinger equation.

D. Interpretation of Schrödinger Equation Results

The idea underlying the Schrödinger equation is that the wavelike behavior of matter, as postulated by de Broglie in 1923 and confirmed by the electron diffraction experiments of Davisson and Germer and by G. P. Thomson in 1927, should be capable of being described within the mathematical framework of a wave equation. The concept of wave equations, of course, underlies all of the present work, but Schrödinger was the first to apply this idea to a theory for describing particles, from which he developed that ubiquitous workhorse of present-day quantum mechanics known as the Schrödinger equation.

Now it should be pointed out straightway that the Schrödinger equation was not in any strict sense "derived" by Schrödinger. In fact, it cannot be rigorously derived. It only can be rationalized in its form and then shown to predict results that agree with experiment. The conventional approach to this rationalization process can be found in nearly any standard quantum mechanics textbook. In the present work, a more novel approach has been developed based on the more general concept of dispersion relations. In many respects, this represents an intuitively satisfying approach. Be that as it may, the time-independent and time-dependent Schrödinger equations are usually written in the forms given by Eqs. (169) and (179), respectively. The complexity of the differential equations given by the time-dependent and time-independent Schrödinger equations is related directly to the complexity of the potential energy $U(\mathbf{r})$ for the problem in question.

Solutoin of the Schrödinger equation gives functions $\psi^{(\mathscr{S}ch)}$ that are usually complex. The square of the magnitude of $\psi^{(\mathscr{S}ch)}$ evaluated at position $\mathbf{r}$ has been interpreted by Born (1882–1970) as the relative probability that the particle is located at that position. By judicious choice of values for the otherwise unspecified constants, such as T_0 and θ_0 in Eqs. (156) and (172), respectively, the integral of the probability over all space can be set equal to unity, a process called normalization. This is physically meaningful because the particle is presumed to be somewhere but not to be at more than one place at a given time. The normalization process requires that the wavefunction decrease sufficiently rapidly with distance away from the general location of the particle to give a bounded value for the integral of the square of the function. This restriction severely limits the physical set of solutions from the great number of mathematical solutions that formally satisfy the Schrödinger equation.

For the hydrogen atom, for example, where the relevant potential energy for the time-independent Schrödinger equation [Eq. (169)] is given as $U(\mathbf{r}) = -e^2/4\pi\varepsilon_0 r$ by Eqs. (2) and (5), with $Z = 1$, the various functions ϕ_j satisfying the equation and meeting the physically reasonable boundary conditions constitute a discrete set. The attendant energies $\mathscr{E}_j$ are essentially those given by Eq. (33), which are deduced by the simpler Bohr theory. A logical extension of the potential energy to include the energy of the electron-spin magnetic moment in an external magnetic field leads to a close correlation of the quantum mechanical predictions for the energy levels of the one-electron atom with detailed, experimental, optical spectral data.

Other standard problems can be attacked with the knowledge developed to the present point. One such problem is that of the one-dimensional, simple harmonic oscillator, characterized by the potential energy $U(x) = \frac{1}{2}Kx^2$, where K is a constant. This problem has applications in a number of fields; e.g., the harmonic oscillator is important for quantized lattice vibrations (phonons) in solids.

Other revealing problems are those of a particle trapped in one-dimensional and three-dimensional boxes. The "particle in a box" model is quite important for the free-electron theory of metals as well as for the consideration of present-day quantum-well semiconductor devices. The general problem is actually quite analogous to the analysis of electromagnetic radiation at thermal equilibrium with the conducting walls of a cavity. The density of allowed modes is computed essentially the same way for both problems.

Later, several standard one-electron problems will be examined. However, let us first develop the concept of quantum mechanical current density and apply it to problems involving constant potentials for which plane-wave solutions to the Schrödinger equation are appropriate.

VII. QUANTUM MECHANICAL CURRENT DENSITY AND PARTICLE BEAMS

A. Probability Current Density

The probability density $\rho = |\psi(\mathbf{r}, t)|^2$ represents a convenient starting point for the consideration of particle and charge currents as computed in quantum mechanics. If this is considered to be a statistical quantity that is a continuous function of position, then the time derivative gives the rate of change of the particle density with time

$$\begin{aligned}\frac{\partial\rho}{\partial t} &= \frac{\partial}{\partial t}|\psi(\mathbf{r}, t)|^2 = \frac{\partial}{\partial t}[\psi^*(\mathbf{r}, t)\psi(\mathbf{r}, t)]\\ &= \psi^*\frac{\partial\psi}{\partial t} + \frac{\partial\psi^*}{\partial t}\psi. \qquad (180)\end{aligned}$$

However, a time rate of change of the probability density at any given point in space requires a difference between the particle currents flowing into and out of the differential volume surrounding the point in question. The mathematical statement of this fact is the well-known microscopic equation of continuity

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot \mathbf{J} \tag{181}$$

where $\nabla \cdot \mathbf{J}$ is the divergence of the particle current $\mathbf{J}$ at the point in question.

Equating Eqs. (180) and (181) for $\partial\rho/\partial t$ gives the relation

$$\nabla \cdot \mathbf{J} = -\left(\psi^* \frac{\partial \psi}{\partial t} + \frac{\partial \psi^*}{\partial t}\psi\right) \tag{182}$$

that must be obeyed by the quantum mechanical analog of the particle current density $\mathbf{J}$. For further development of this expression, the time-dependent Schrödinger equation (179) and its complex conjugate can be employed:

$$i\hbar \frac{\partial \psi}{\partial t} = -\frac{\hbar^2}{2m}\nabla^2\psi + U(\mathbf{r})\psi \tag{183}$$

$$-i\hbar \frac{\partial \psi^*}{\partial t} = -\frac{\hbar^2}{2m}\nabla^2\psi^* + U(\mathbf{r})\psi^*. \tag{184}$$

Taking the complex conjugate utilizes the relations $\mathbf{r}^* = \mathbf{r}$, $t^* = t$, and $U(\mathbf{r})^* = U(\mathbf{r})$ due to the fact that the concern is with real positions, real times, and real potential energies. Multiplying the Schrödinger equation by ψ^* and its complex conjugate by ψ gives

$$i\hbar\psi^* \frac{\partial \psi}{\partial t} = -\frac{\hbar^2}{2m}\psi^*\nabla^2\psi + \psi^* U(\mathbf{r})\psi \tag{185}$$

$$-i\hbar\psi \frac{\partial \psi^*}{\partial t} = -\frac{\hbar^2}{2m}\psi\nabla^2\psi^* + \psi U(\mathbf{r})\psi^* \tag{186}$$

Subtracting Eq. (186) from Eq. (185) gives

$$i\hbar\left(\psi^* \frac{\partial \psi}{\partial t} + \psi \frac{\partial \psi^*}{\partial t}\right) = \frac{\hbar^2}{2m}(\psi\nabla^2\psi^* - \psi^*\nabla^2\psi), \tag{187}$$

where, in equating $\psi^* U(\mathbf{r})\psi$ with $\psi U(\mathbf{r})\psi^*$, the property of $U(\mathbf{r})$ is utilized merely as a multiplicative operator, so that the factors in the product $\psi^* U(\mathbf{r})\psi$ commute. The right-hand side of Eq. (187) involves the Laplacian operator, so that it is reasonable to expect that perhaps it can be expressed as the divergence of some vector quantity. Taking the divergence of $\psi^*\nabla\psi$ yields

$$\nabla \cdot (\psi^*\nabla\psi) = \nabla\psi^* \cdot \nabla\psi + \psi^*\nabla^2\psi, \tag{188}$$

whereas taking the divergence of $\psi\nabla\psi^*$ gives

$$\nabla \cdot (\psi\nabla\psi^*) = \nabla\psi \cdot \nabla\psi^* + \psi\nabla^2\psi^*. \tag{189}$$

Recognizing that the dot product of two vectors such as $\nabla\psi \cdot \nabla\psi^*$ is commutative, so that $\nabla\psi \cdot \nabla\psi^* = \nabla\psi^* \cdot \nabla\psi$, these two quantities can be subtracted to give the relation

$$\nabla \cdot (\psi^*\nabla\psi - \psi\nabla\psi^*) = \psi^*\nabla^2\psi - \psi\nabla^2\psi^*. \tag{190}$$

The right-hand side of Eq. (190) can be identified with the factor in the right-hand side of Eq. (187), so that

$$i\hbar\left(\psi^* \frac{\partial \psi}{\partial t} + \psi \frac{\partial \psi^*}{\partial t}\right) = -\frac{\hbar^2}{2m}\nabla \cdot (\psi^*\nabla\psi - \psi\nabla\psi^*). \tag{191}$$

Substituting into Eq. (182) for $\nabla \cdot \mathbf{J}$ gives

$$\nabla \cdot \mathbf{J} = \frac{\hbar}{2mi}\nabla \cdot (\psi^*\nabla\psi - \psi\nabla\psi^*). \tag{192}$$

Within an arbitrary constant, then

$$\mathbf{J} = \frac{\hbar}{2mi}(\psi^*\nabla\psi - \psi\nabla\psi^*). \tag{193}$$

The arbitrary constant is zero if $\mathbf{J} = 0$ whenever $\psi = 0$, as one would expect. Knowledge of the wavefunction ψ therefore allows one to calculate the particle current density $\mathbf{J}$ in quantum mechanics. For the case of electrons, the charge per particle is $-e$, so that the charge density $\mathscr{J}$ follows immediately from

$$\mathscr{J} = -e\mathbf{J}. \tag{194}$$

It is illuminating to apply the relation given by Eq. (193) to the specific case of plane waves $\exp(i\mathbf{k} \cdot \mathbf{r})$, which can be shown to be eigenfunctions of the so-called momentum operator $\mathfrak{p}^{\mathfrak{op}} = -i\hbar\nabla$. Thus, $\nabla\psi = (i/\hbar)\mathbf{p}\psi$. Also $\nabla\psi^* = (-i/\hbar)\mathbf{p}\psi^*$, so that

$$\begin{aligned}\mathbf{J} &= \frac{\hbar}{2mi}\left(\psi^* \frac{i}{\hbar}\mathbf{p}\psi - \psi\frac{(-i)}{\hbar}\mathbf{p}\psi^*\right)\\ &= \frac{\mathbf{p}}{2m}(\psi^*\psi + \psi^*\psi) = \psi^*\psi\mathbf{v}.\end{aligned} \tag{195}$$

This is simply the product of the particle probability density $\psi^*\psi$ and the particle velocity $\mathbf{v}$, which is readily interpreted as the particle current density on the basis of physical considerations.

B. Piecewise Constant Potential Energy Problems

It is worthwhile to work through the details of an illustrative example that typifies the quantum treatment of a particle-beam incident on potential energy steps, rectangular potential barriers and wells, and arrays of configurations that may be useful in understanding currents in modern solid-state devices. Figure 4 illustrates the three regions defined by the potential energy function chosen for this example:

$$U(x) = \begin{cases} 0 & (x < 0) \\ U_0 & (0 \le x \le L). \\ W & (x > L) \end{cases} \tag{196}$$

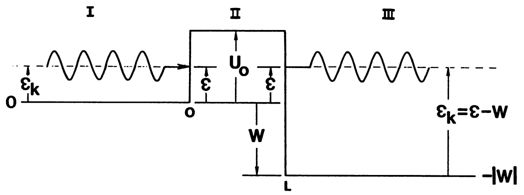


FIGURE 4 Piecewise constant potential energy regions. [Fig. 1.34 in *Quantum Mechanics for Applied Physics and Engineering* by Albert Thomas Fromhold, Jr. (Academic Press, Inc., New York, 1981; Dover Publications, Inc., New York, 1991); reproduced with the permission of Academic Press, Dover Publications, and the author.]

The figure is drawn with $U_0 > 0$ and $W < 0$, but actually either sign is possible for either parameter. Whenever $W \rightarrow U_0$, the limit is a single potential step; it constitutes a step up for $U_0 > 0$ but a step down for $U_0 < 0$. Whenever $W \rightarrow 0$, the limit is a rectangular barrier if $U_0 > 0$ but a rectangular potential well if $U_0 < 0$. The rectangular potential barrier is a common example used to illustrate quantum tunneling, which is a penetration of the barrier for particle energies $\mathscr{E}$ below the barrier height U_0, even though such penetration would be disallowed from the standpoint of classical physics. The results are immediately applicable to the tunneling of electrons between metals separated by a thin insulator.

Before considering the several possible cases individually, let us first clarify the predictions of classical physics for this example to sharpen our understanding of the problem and to highlight the differences in the predictions of the quantum mechanics and classical mechanics theories. Classically, if $\mathscr{E} > U_0$ and $\mathscr{E} > W$, there is total transmission, since the momentum of the particles does not change sign and is not decreased to zero as the particle crosses the barriers. On the other hand, the classical picture states that whenever $\mathscr{E} < U_0$, all particles will be reflected by the barrier, so there should be no transmission through the barrier.

As will now be shown, the quantum mechanics predictions are somewhat different. The classical mechanics result is more straightforward in a sense, because the reflection or transmission, as the case may be, is total. The quantum mechanics results, on the other hand, are a bit more mysterious, there being cases of partial transmission and partial reflection of a particle beam. The quantum mechanics predictions, however, are found to agree with the experimental results.

C. Incident Beam with Particle Energy Exceeding Both Steps

This is the case for $\mathscr{E} > U_0$ and $\mathscr{E} > W$, algebraically speaking. All wavefunctions will be of the plane-wave type. The momentum is real, and propagation of the particle is possible, even in the classical sense. The problem will be developed in terms of the Schrödinger picture, using ψ and ϕ, although the problem can be developed equally well within the framework of a relativistic master equation given by Eq. (166), using Ψ and Φ. Let the incident wave be given by

$$\psi_{\text{inc}} = Ae^{i(kx-\omega t)} \tag{197}$$

where, in the Schrödinger picture

$$k = \hbar^{-1}(2m\mathscr{E})^{1/2} \tag{198}$$

with $\mathscr{E}$ representing the total Newtonian energy $\mathscr{E}_{\mathrm{T}}^{(\mathcal{N}t)}$ and with

$$\omega = \frac{\mathscr{E}}{\hbar} \tag{199}$$

for this picture, in contrast to the ω given by Eq. (150) in the relativistic picture. The corresponding reflected wave can be written as

$$\psi_{\text{ref}} = Be^{i(-kx-\omega t)} \tag{200}$$

Then for region I in Fig. 4

$$\psi_{\text{I}} = \psi_{\text{inc}} + \psi_{\text{ref}} = (Ae^{ikx} + Be^{-ikx})e^{-i\omega t} \tag{201}$$

The transmitted wave is the propagating wave in region III. Let us denote the transmitted wave by

$$\psi_{\text{trans}} = Ce^{i(\mathscr{k}x-\omega t)}, \tag{202}$$

where

$$\mathscr{k} = \hbar^{-1}[2m(\mathscr{E} - W)]^{1/2}. \tag{203}$$

In the absence of sources and other variations in region III that could lead to a reflected wave there, then

$$\psi_{\text{III}} = \psi_{\text{trans}} = Ce^{i\mathscr{k}x}e^{-i\omega t}. \tag{204}$$

Region II must now be considered. Due to the finite thickness of the region $(0 \le x \le L)$ and the discontinuity at $x = L$, it is possible to have a reverse traveling (reflected) wave in this region in addition to a forward propagating wave. Denote the forward wave by $[Fe^{i(\beta x-\omega t)}]$ and the reverse wave by $[Ge^{i(-\beta x-\omega t)}]$, where

$$\beta \equiv \hbar^{-1}[2m(\mathscr{E} - U_0)]^{1/2} \tag{205}$$

then

$$\psi_{\text{II}} = (Fe^{i\beta x} + Ge^{-i\beta x})e^{-i\omega t}. \tag{206}$$

The boundary condition at $x = 0$ of wavefunction continuity

$$\psi_{\text{I}}(0) = \psi_{\text{II}}(0) \tag{207}$$

is sufficient to ensure continuity of the particle density. The boundary condition of continuity of the first derivative of the wavefunction

$$\left.\frac{d\psi_{\text{I}}}{dx}\right|_{x=0} = \left.\frac{d\psi_{\text{II}}}{dx}\right|_{x=0} \tag{208}$$

is sufficient to ensure continuity of the current density, as can be noted from Eq. (193). These two conditions lead directly to the following two relations:

$$A + B = F + G \tag{209}$$

$$ikA - ikB = i\beta F - i\beta G. \tag{210}$$

Rewriting this pair of equations in the form

$$A + B = F + G \tag{211}$$

$$A - B = \left(\frac{\beta}{k}\right)\{F - G\} \tag{212}$$

makes it easy to obtain expressions for A and B in terms of F and G. Adding the two equations gives

$$A = \frac{1}{2}\left[F\left(1 + \frac{\beta}{k}\right) + G\left(1 - \frac{\beta}{k}\right)\right] \tag{213}$$

and subtracting the two equations gives

$$B = \frac{1}{2}\left[F\left(1 - \frac{\beta}{k}\right) + G\left(1 + \frac{\beta}{k}\right)\right]. \tag{214}$$

Let us next apply boundary conditions at the barrier discontinuity at $x = L$. Continuity of the wavefunction $\psi_{\text{II}}(L) = \psi_{\text{III}}(L)$ and continuity of the first derivative of the wavefunction

$$\left.\frac{d\psi_{\text{II}}}{dx}\right|_{x=L} = \left.\frac{d\psi_{\text{III}}}{dx}\right|_{x=L}$$

lead directly to two additional relations:

$$Fe^{i\beta L} + Ge^{-i\beta L} = Ce^{i\Bbbk L} \tag{215}$$

$$i\beta Fe^{i\beta L} - i\beta Ge^{-i\beta L} = i\Bbbk Ce^{i\Bbbk L}. \tag{216}$$

Rewriting this pair of equations in the form

$$Fe^{i\beta L} + Ge^{-i\beta L} = Ce^{i\Bbbk L} \tag{217}$$

$$Fe^{i\beta L} - Ge^{-i\beta L} = \frac{\Bbbk}{\beta}Ce^{i\Bbbk L} \tag{218}$$

leads to expressions for F and G in terms of C. Adding the two equations gives

$$\begin{aligned} F &= \frac{1}{2}e^{-i\beta L}\left(1 + \frac{\Bbbk}{\beta}\right)Ce^{i\Bbbk L} \\ &= \frac{1}{2}\left(1 + \frac{\Bbbk}{\beta}\right)Ce^{i(\Bbbk - \beta)L}. \end{aligned} \tag{219}$$

and subtracting the two equations gives

$$\begin{aligned} G &= \frac{1}{2}e^{i\beta L}\left(1 - \frac{\Bbbk}{\beta}\right)Ce^{i\Bbbk L} \\ &= \frac{1}{2}\left(1 - \frac{\Bbbk}{\beta}\right)Ce^{i(\Bbbk + \beta)L}. \end{aligned} \tag{220}$$

Substituting the two expressions just obtained for F and G into the expressions previously obtained for A and B gives A and B in terms of C:

$$\begin{aligned} A &= \frac{1}{2}\left\{\left(1 + \frac{\beta}{k}\right)\left[\frac{1}{2}\left(1 + \frac{\Bbbk}{\beta}\right)Ce^{i(\Bbbk - \beta)L}\right]\right. \\ &\quad \left. + \left(1 - \frac{\beta}{k}\right)\left[\frac{1}{2}\left(1 - \frac{\Bbbk}{\beta}\right)Ce^{i(\Bbbk + \beta)L}\right]\right\} \\ &= \frac{1}{4}\left[\left(1 + \frac{\beta}{k}\right)\left(1 + \frac{\Bbbk}{\beta}\right)e^{-i\beta L}\right. \\ &\quad \left. + \left(1 - \frac{\beta}{k}\right)\left(1 - \frac{\Bbbk}{\beta}\right)e^{i\beta L}\right]Ce^{i\Bbbk L}. \end{aligned} \tag{221}$$

$$\begin{aligned} B &= \frac{1}{2}\left\{\left(1 - \frac{\beta}{k}\right)\left[\frac{1}{2}\left(1 + \frac{\Bbbk}{\beta}\right)Ce^{i(\Bbbk - \beta)L}\right]\right. \\ &\quad \left. + \left(1 + \frac{\beta}{k}\right)\left[\frac{1}{2}\left(1 - \frac{\Bbbk}{\beta}\right)Ce^{i(\Bbbk + \beta)L}\right]\right\} \\ &= \frac{1}{4}\left[\left(1 - \frac{\beta}{k}\right)\left(1 + \frac{\Bbbk}{\beta}\right)e^{-i\beta L}\right. \\ &\quad \left. + \left(1 + \frac{\beta}{k}\right)\left(1 - \frac{\Bbbk}{\beta}\right)e^{i\beta L}\right]Ce^{i\Bbbk L}. \end{aligned} \tag{222}$$

Thus, the relationship between A and C and the relationship between B and C are obtained.

The product $[(A/C)^*(A/C)]$, which is useful for the transmission coefficient, can now be evaluated:

$$\begin{aligned} \left(\frac{A}{C}\right)^*\left(\frac{A}{C}\right) &= \frac{1}{16}\left\{\left[\left(1 + \frac{\beta}{k}\right)\left(1 + \frac{\Bbbk}{\beta}\right)e^{i\beta L}\right.\right. \\ &\quad \left. + \left(1 - \frac{\beta}{k}\right)\left(1 - \frac{\Bbbk}{\beta}\right)e^{-i\beta L}\right] \\ &\quad \times \left[\left(1 + \frac{\beta}{k}\right)\left(1 + \frac{\Bbbk}{\beta}\right)e^{-i\beta L}\right. \\ &\quad \left.\left. + \left(1 - \frac{\beta}{k}\right)\left(1 - \frac{\Bbbk}{\beta}\right)e^{i\beta L}\right]\right\} \\ &= \frac{1}{16}\left\{\left(1 + \frac{\beta}{k}\right)^2\left(1 + \frac{\Bbbk}{\beta}\right)^2\right. \\ &\quad + \left(1 - \frac{\beta}{k}\right)^2\left(1 - \frac{\Bbbk}{\beta}\right)^2 \\ &\quad + e^{2i\beta L}\left[1 - \left(\frac{\beta}{k}\right)^2\right]\left[1 - \left(\frac{\Bbbk}{\beta}\right)^2\right] \\ &\quad \left. + e^{-2i\beta L}\left[1 - \left(\frac{\beta}{k}\right)^2\right]\left[1 - \left(\frac{\Bbbk}{\beta}\right)^2\right]\right\} \\ &= \frac{1}{16}\left\{\left(1 + \frac{\Bbbk}{k} + \frac{\beta}{k} + \frac{\Bbbk}{\beta}\right)^2\right. \\ &\quad + \left(1 + \frac{\Bbbk}{k} - \frac{\beta}{k} - \frac{\Bbbk}{\beta}\right)^2 \end{aligned}$$

$$+ 2(\cos 2\beta L)\left[1 - \left(\frac{\mathscr{k}}{\beta}\right)^2 - \left(\frac{\beta}{k}\right)^2 + \left(\frac{\mathscr{k}}{k}\right)^2\right]\Bigg\}. \tag{223}$$

The transmission coefficient $\mathcal{T}$ follows from the ratio of the transmitted intensity I_{trans} to the incident-beam intensity I_{inc}:

$$\mathcal{T} = \frac{I_{\text{trans}}}{I_{\text{inc}}} = \frac{\psi^*_{\text{trans}}\psi_{\text{trans}}(\hbar\mathscr{k}/m)}{\psi^*_{\text{inc}}\psi_{\text{inc}}(\hbar k/m)} = \left(\frac{C^*C}{A^*A}\right)\left(\frac{\mathscr{k}}{k}\right) = \frac{\mathscr{k}/k}{\left[\left(\frac{A}{C}\right)^*\left(\frac{A}{C}\right)\right]}. \tag{224}$$

Substituting the evaluation for the denominator given in Eq. (223) then yields the transmission coefficient.

The reflection coefficient $\mathcal{R}$ can be obtained from the ratio of the reflected intensity to the incident-beam intensity:

$$\mathcal{R} = \left|\frac{I_{\text{ref}}}{I_{\text{inc}}}\right| = \left|\frac{B^*B(-\hbar k/m)}{A^*A(\hbar k/m)}\right| = \left(\frac{B}{A}\right)^*\left(\frac{B}{A}\right). \tag{225}$$

Already B in terms of C and A in terms of C have been obtained. The ratio of these two expressions then gives

$$\frac{B}{A} = \frac{[1-(\beta/k)][1+(\mathscr{k}/\beta)]e^{-i\beta L} + [1+(\beta/k)][1-(\mathscr{k}/\beta)]e^{i\beta L}}{[1+(\beta/k)][1+(\mathscr{k}/\beta)]e^{-i\beta L} + [1-(\beta/k)][1-(\mathscr{k}/\beta)]e^{i\beta L}}. \tag{226}$$

Thus

$$\mathcal{R} = \left(\frac{B}{A}\right)^*\left(\frac{B}{A}\right)$$
$$= \frac{\left[\left\{\left(1-\frac{\beta}{k}\right)^2\left(1+\frac{\mathscr{k}}{\beta}\right)^2 + \left(1+\frac{\beta}{k}\right)^2\left(1-\frac{\mathscr{k}}{\beta}\right)^2 + e^{-2i\beta L}\left[1-\left(\frac{\beta}{k}\right)^2\right]\left[1-\left(\frac{\mathscr{k}}{\beta}\right)^2\right] + e^{2i\beta L}\left[1-\left(\frac{\beta}{k}\right)^2\right]\left[1-\left(\frac{\mathscr{k}}{\beta}\right)^2\right]\right\}\right]}{\left[\left\{\left(1+\frac{\beta}{k}\right)^2\left(1+\frac{\mathscr{k}}{\beta}\right)^2 + \left(1-\frac{\beta}{k}\right)^2\left(1-\frac{\mathscr{k}}{\beta}\right)^2 + e^{-2i\beta L}\left[1-\left(\frac{\beta}{k}\right)^2\right]\left[1-\left(\frac{\mathscr{k}}{\beta}\right)^2\right] + e^{2i\beta L}\left[1-\left(\frac{\beta}{k}\right)^2\right]\left[1-\left(\frac{\mathscr{k}}{\beta}\right)^2\right]\right\}\right]} \tag{227}$$

or, equivalently

$$\mathcal{R} = \frac{\{[1-(\beta/k)+(\mathscr{k}/\beta)-(\mathscr{k}/k)]^2 + [1+(\beta/k)-(\mathscr{k}/\beta)-(\mathscr{k}/k)]^2 + 2\{\cos 2\beta L\}[1-(\beta/k)^2-(\mathscr{k}/\beta)^2+(\mathscr{k}/k)^2]\}}{\{[1-(\beta/k)+(\mathscr{k}/\beta)+(\mathscr{k}/k)]^2 + [1-(\beta/k)-(\mathscr{k}/\beta)+(\mathscr{k}/k)]^2 + 2\{\cos 2\beta L\}[1-(\beta/k)^2-(\mathscr{k}/\beta)^2+(\mathscr{k}/k)^2]\}} \tag{228}$$

It can be shown that $\mathcal{R} = (1 - \mathcal{T})$, which means that whatever current density is not transmitted is reflected.

Now let us consider the limit $W \to 0$, in which case $\mathscr{k} \to k$ and

$$\begin{aligned}\left(\frac{A}{C}\right)^*\left(\frac{A}{C}\right) &= \frac{1}{16}\Bigg\{\left[2+\left(\frac{\beta}{k}+\frac{k}{\beta}\right)\right]^2 \\ &\quad + \left[2-\left(\frac{\beta}{k}+\frac{k}{\beta}\right)\right]^2 \\ &\quad + 2(\cos 2\beta L)\left[2-\left(\frac{k}{\beta}\right)^2-\left(\frac{\beta}{k}\right)^2\right]\Bigg\} \\ &= \frac{1}{16}\Bigg\{2\left[4+\left(\frac{\beta}{k}+\frac{k}{\beta}\right)^2\right] \\ &\quad + 2(\cos 2\beta L)\left[2-\left(\frac{k}{\beta}\right)^2-\left(\frac{\beta}{k}\right)^2\right]\Bigg\} \\ &= \frac{1}{8}\Bigg\{\left[6+\left(\frac{\beta}{k}\right)^2+\left(\frac{k}{\beta}\right)^2\right] \\ &\quad + \left[2-\left(\frac{\beta}{k}\right)^2-\left(\frac{k}{\beta}\right)^2\right]\cos 2\beta L\Bigg\}.\end{aligned} \tag{229}$$

Thus, the following limit is obtained:

$$\mathcal{T} = \frac{8}{[6+(\beta/k)^2+(k/\beta)^2]+[2-(\beta/k)^2-(k/\beta)^2]\cos 2\beta L}. \tag{230}$$

This rectangular barrier solution contrasts markedly with the classical result $\mathscr{T} = 1$. The cosine factor in the denominator leads to a decided oscillatory dependence of the transmission coefficient on energy of the incident particles, as will be demonstrated later in a graphical example. If, in addition, the barrier height U_0 is taken to be zero, then $\beta = k$ and $\mathscr{T} = 1$, as expected. This limit is also well approximated for $\mathscr{E} \gg U_0$, since then $\beta \simeq k$.

In the alternate limit, $W \to U_0$, so that $\mathscr{k} \to \beta$, and Eq. (224) yields

$$\mathscr{T} = \frac{4\mathscr{k}k}{(k + \mathscr{k})^2}. \tag{231}$$

This is the transmission coefficient for a step potential of height U_0 for the case $\mathscr{E} > U_0$. In this same limit $W \to U_0$, Eq. (228) for the reflection coefficient reduces to

$$\mathscr{R} = \frac{(k - \mathscr{k})^2}{(k + \mathscr{k})^2}. \tag{232}$$

D. Incident Beam with Particle Energy below First Step Only

Let us now consider a case for which $W < \mathscr{E} < U_0$ algebraically. For $\mathscr{E} < U_0$, the wavefunction in region II (see Fig. 4) cannot be of the plane-wave type, because the kinetic energy would be negative and the momentum consequently would be imaginary. It is easy to show that the wavefunction

$$\psi_{\mathrm{II}} = (De^{-\alpha x} + Ee^{\alpha x})e^{-i\omega t} \tag{233}$$

satisfies the Schrödinger equation

$$\left(-\frac{\hbar^2}{2m}\right)\left(\frac{\partial^2\psi}{\partial x^2}\right) + U_0\psi = i\hbar\frac{\partial\psi}{\partial t} \tag{234}$$

appropriate for this region, where α is obtained by substituting ψ_{II} into the equation

$$-\frac{\hbar^2}{2m}\alpha^2\psi_{\mathrm{II}} + U_0\psi_{\mathrm{II}} = i\hbar(-i\omega)\psi_{\mathrm{II}}, \tag{235}$$

which gives

$$-\frac{\hbar^2}{2m}\alpha^2 = \hbar\omega - U_0 \tag{236}$$

or, since $\hbar\omega = \mathscr{E}$

$$\alpha = \left[\frac{2m}{\hbar^2}(U_0 - \mathscr{E})\right]^{1/2}. \tag{237}$$

The positive sign is conventionally chosen for α; the choice of a negative sign would simply interchange coefficients D and E.

Since $\mathscr{E} > W$, the solution in region III (see Fig. 4) is again of the propagating type:

$$\psi_{\mathrm{III}} = \psi_{\mathrm{trans}} = Ce^{i(\mathscr{k}x - \omega t)} = Ce^{i\mathscr{k}x}e^{-i\omega t}. \tag{238}$$

As before, the wavefunction ψ_1 for region I is

$$\begin{aligned}\psi_{\mathrm{I}} &= \psi_{\mathrm{inc}} + \psi_{\mathrm{ref}} \\ &= (Ae^{ikx} + Be^{-ikx})e^{-i\omega t},\end{aligned} \tag{239}$$

where the constants $\mathscr{k}$ and k have their previously defined values

$$\mathscr{k} = \hbar^{-1}[2m(\mathscr{E} - W)]^{1/2} \tag{240}$$

$$k = \hbar^{-1}(2m\mathscr{E})^{1/2}. \tag{241}$$

The boundary conditions at $x = 0$ of continuity of the wavefunction $\psi_{\mathrm{I}}(0) = \psi_{\mathrm{II}}(0)$ and continuity of the first derivative of the wavefunction

$$\left.\frac{d\psi_{\mathrm{I}}}{dx}\right|_{x=0} = \left.\frac{d\psi_{\mathrm{II}}}{dx}\right|_{x=0}$$

lead directly to the following two relations:

$$A + B = D + E \tag{242}$$

$$ikA - ikB = -\alpha D + \alpha E. \tag{243}$$

Rewriting this pair of equations in the form

$$A + B = D + E \tag{244}$$

$$A - B = \frac{-\alpha}{ik}(D - E) \tag{245}$$

makes it easy to obtain expressions for A and B in terms of D and E. Adding the two equations gives

$$A = \frac{1}{2}\left[D\left(1 - \frac{\alpha}{ik}\right) + E\left(1 + \frac{\alpha}{ik}\right)\right] \tag{246}$$

and subtracting the two equations gives

$$B = \frac{1}{2}\left[D\left(1 + \frac{\alpha}{ik}\right) + E\left(1 - \frac{\alpha}{ik}\right)\right]. \tag{247}$$

Next, let us apply boundary conditions at the barrier discontinuity at $x = L$. The conditions of continuity of the wavefunction $\psi_{\mathrm{II}}(L) = \psi_{\mathrm{III}}(L)$ and continuity of the first derivative of the wavefunction

$$\left.\frac{d\psi_{\mathrm{II}}}{dx}\right|_{x=L} = \left.\frac{d\psi_{\mathrm{III}}}{dx}\right|_{x=L}$$

lead directly to the two additional relations:

$$De^{-\alpha L} + Ee^{\alpha L} = Ce^{i\mathscr{k}L} \tag{248}$$

$$-\alpha De^{-\alpha L} + \alpha Ee^{\alpha L} = i\mathscr{k}Ce^{i\mathscr{k}L}. \tag{249}$$

Rewriting this pair of equations in the form

$$De^{-\alpha L} + Ee^{\alpha L} = Ce^{i\mathscr{k}L} \tag{250}$$

$$De^{-\alpha L} - Ee^{\alpha L} = \frac{-i\mathscr{k}}{\alpha}Ce^{i\mathscr{k}L}. \tag{251}$$

leads to expressions for D and E in terms of C. Adding the two equations gives

$$D = \frac{1}{2}e^{\alpha L}\left(1 + \frac{-i\Bbbk}{\alpha}\right)Ce^{i\Bbbk L}$$
$$= \frac{1}{2}\left(1 - \frac{i\Bbbk}{\alpha}\right)Ce^{(i\Bbbk+\alpha)L} \qquad (252)$$

and subtracting the two equations gives

$$E = \frac{1}{2}e^{-\alpha L}\left(1 + \frac{i\Bbbk}{\alpha}\right)Ce^{i\Bbbk L}$$
$$= \frac{1}{2}\left(1 + \frac{i\Bbbk}{\alpha}\right)Ce^{(i\Bbbk-\alpha)L}. \qquad (253)$$

Substituting these two expressions for D and E into the expressions obtained in Eqs. (246) and (247) for A and B gives A and B in terms of C:

$$A = \frac{1}{2}\left\{\left(1 - \frac{\alpha}{ik}\right)\left[\frac{1}{2}\left(1 - \frac{i\Bbbk}{\alpha}\right)Ce^{(i\Bbbk+\alpha)L}\right] + \left(1 + \frac{\alpha}{ik}\right)\left[\frac{1}{2}\left(1 + \frac{i\Bbbk}{\alpha}\right)Ce^{(i\Bbbk-\alpha)L}\right]\right\}$$
$$= \frac{1}{4}\left[\left(1 - \frac{\alpha}{ik}\right)\left(1 - \frac{i\Bbbk}{\alpha}\right)e^{\alpha L} + \left(1 + \frac{\alpha}{ik}\right)\left(1 + \frac{i\Bbbk}{\alpha}\right)e^{-\alpha L}\right]Ce^{i\Bbbk L}. \qquad (254)$$

$$B = \frac{1}{2}\left\{\left(1 + \frac{\alpha}{ik}\right)\left[\frac{1}{2}\left(1 - \frac{i\Bbbk}{\alpha}\right)Ce^{(i\Bbbk+\alpha)L}\right] + \left(1 - \frac{\alpha}{ik}\right)\left[\frac{1}{2}\left(1 + \frac{i\Bbbk}{\alpha}\right)Ce^{(i\Bbbk-\alpha)L}\right]\right\}$$
$$= \frac{1}{4}\left[\left(1 + \frac{\alpha}{ik}\right)\left(1 - \frac{i\Bbbk}{\alpha}\right)e^{\alpha L} + \left(1 - \frac{\alpha}{ik}\right)\left(1 + \frac{i\Bbbk}{\alpha}\right)e^{-\alpha L}\right]Ce^{i\Bbbk L}. \qquad (255)$$

Thus, the relationships between A and C and between B and C have been obtained. The ratio of B to A now is easily obtained:

$$\frac{B}{A} = \frac{[1 + (\alpha/ik)][1 - (i\Bbbk/\alpha)]e^{\alpha L} + [1 + (\alpha/ik)][1 + (i\Bbbk/\alpha)]e^{-\alpha L}}{[1 - (\alpha/ik)][1 - (i\Bbbk/\alpha)]e^{\alpha L} + [1 + (\alpha/ik)][1 + (i\Bbbk/\alpha)]e^{-\alpha L}}$$
$$= \frac{[1 - (i\Bbbk/\alpha) + (\alpha/ik) - (\Bbbk/k)]e^{\alpha L} + [1 + (i\Bbbk/\alpha) - (\alpha/ik) - (\Bbbk/k)]e^{-\alpha L}}{[1 - (i\Bbbk/\alpha) - (\alpha/ik) + (\Bbbk/k)]e^{\alpha L} + [1 + (i\Bbbk/\alpha) + (\alpha/ik) + (\Bbbk/k)]e^{-\alpha L}}$$
$$= \frac{\{[1 - (\Bbbk/k)] - i[(\Bbbk/\alpha) + (\alpha/k)]\}e^{\alpha L} + \{[1 - (\Bbbk/k)] + i[(\Bbbk/\alpha) + (\alpha/k)]\}e^{-\alpha L}}{\{[1 + (\Bbbk/k)] - i[(\Bbbk/\alpha) - (\alpha/k)]\}e^{\alpha L} + \{[1 - (\Bbbk/k)] + i[(\Bbbk/\alpha) - (\alpha/k)]\}e^{-\alpha L}}$$
$$= \frac{[1 - (\Bbbk/k)](e^{\alpha L} + e^{-\alpha L}) - i[(\Bbbk/\alpha) + (\alpha/k)](e^{\alpha L} - e^{-\alpha L})}{[1 + (\Bbbk/k)](e^{\alpha L} + e^{-\alpha L}) - i[(\Bbbk/\alpha) - (\alpha/k)](e^{\alpha L} - e^{-\alpha L})}$$
$$= \frac{[1 - (\Bbbk/k)]\cosh(\alpha L) - i[(\Bbbk/\alpha) + (\alpha/k)]\sinh(\alpha L)}{[1 + (\Bbbk/k)]\cosh(\alpha L) - i[(\Bbbk/\alpha) - (\alpha/k)]\sinh(\alpha L)}. \qquad (256)$$

The reflection coefficient $\mathcal{R}$ can be obtained from the ratio of the reflected intensity I_{ref} to the incident-beam intensity I_{inc}:

$$\mathcal{R} = \left|\frac{I_{\text{ref}}}{I_{\text{inc}}}\right| = \left|\frac{B^*B(-\hbar k/m)}{A^*A(\hbar k/m)}\right| = \left(\frac{B}{A}\right)^*\left(\frac{B}{A}\right). \qquad (257)$$

Therefore

$$\mathcal{R} = \frac{[1 - (\Bbbk/k)]^2\cosh^2(\alpha L) + [(\Bbbk/\alpha) + (\alpha/k)]^2\sinh^2(\alpha L)}{[1 + (\Bbbk/k)]^2\cosh^2(\alpha L) + [(\Bbbk/\alpha) - (\alpha/k)]^2\sinh^2(\alpha L)}$$
$$= \frac{[1 - 2(\Bbbk/k) + (\Bbbk/k)^2]\cosh^2(\alpha L) + [(\Bbbk/\alpha)^2 + 2(\Bbbk/k) + (\alpha/k)^2]\sinh^2(\alpha L)}{[1 + 2(\Bbbk/k) + (\Bbbk/k)^2]\cosh^2(\alpha L) + [(\Bbbk/\alpha)^2 - 2(\Bbbk/k) + (\alpha/k)^2]\sinh^2(\alpha L)}$$
$$= \frac{[[\cosh^2(\alpha L) + (\Bbbk/\alpha)^2\sinh^2(\alpha L)] - 2(\Bbbk/k)[\cosh^2(\alpha L) - \sinh^2(\alpha L)] + (\Bbbk/k)^2\cosh^2(\alpha L) + (\alpha/k)^2\sinh^2(\alpha L)]}{[[\cosh^2(\alpha L) + (\Bbbk/\alpha)^2\sinh^2(\alpha L)] + 2(\Bbbk/k)[\cosh^2(\alpha L) - \sinh^2(\alpha L)] + (\Bbbk/k)^2\cosh^2(\alpha L) + (\alpha/k)^2\sinh^2(\alpha L)]}. \qquad (258)$$

Since, for arbitrary θ, $\cosh(\theta) = \frac{1}{2}(e^{\theta} + e^{-\theta})$ and $\sinh(\theta) = \frac{1}{2}(e^{\theta} - e^{-\theta})$, it follows that $\cosh^2(\theta) = 1 + \sinh^2(\theta)$ and $\cosh^2(\theta) + \sinh^2(\theta) = \cosh(2\theta)$. Thus

$$\begin{aligned}
\mathscr{R} &= \frac{1 + [1 + (\mathscr{k}/\alpha)^2]\sinh^2(\alpha L) - 2(\mathscr{k}/k) + (\mathscr{k}/k)^2 + [(\mathscr{k}/k)^2 + (\alpha/k)^2]\sinh^2(\alpha L)}{1 + [1 + (\mathscr{k}/\alpha)^2]\sinh^2(\alpha L) + 2(\mathscr{k}/k) + (\mathscr{k}/k)^2 + [(\mathscr{k}/k)^2 + (\alpha/k)^2]\sinh^2(\alpha L)} \\
&= \frac{[1 - 2(\mathscr{k}/k) + (\mathscr{k}/k)^2] + [1 + (\mathscr{k}/\alpha)^2 + (\mathscr{k}/k)^2 + (\alpha/k)^2]\sinh^2(\alpha L)}{[1 + 2(\mathscr{k}/k) + (\mathscr{k}/k)^2] + [1 + (\mathscr{k}/\alpha)^2 + (\mathscr{k}/k)^2 + (\alpha/k)^2]\sinh^2(\alpha L)} \\
&= \frac{[1 - (\mathscr{k}/k)]^2 + [1 + (\mathscr{k}/\alpha)^2][1 + (\alpha/k)^2]\sinh^2(\alpha L)}{[1 + (\mathscr{k}/k)]^2 + [1 + (\mathscr{k}/\alpha)^2][1 + (\alpha/k)^2]\sinh^2(\alpha L)}. \qquad (259)
\end{aligned}$$

The transmission coefficient $\mathscr{T}$ is readily evaluated from this expression for the reflection coefficient:

$$\begin{aligned}
\mathscr{T} &= 1 - \mathscr{R} \\
&= \frac{[1 + (\mathscr{k}/k)]^2 + [1 + (\mathscr{k}/\alpha)^2][1 + (\alpha/k)^2]\sinh^2(\alpha L) - [1 - (\mathscr{k}/k)]^2 - [1 + (\mathscr{k}/\alpha)^2][1 + (\alpha/k)^2]\sinh^2(\alpha L)}{[1 + (\mathscr{k}/k)]^2 + [1 + (\mathscr{k}/\alpha)^2][1 + (\alpha/k)^2]\sinh^2(\alpha L)} \\
&= \frac{4(\mathscr{k}/k)}{[1 + (\mathscr{k}/k)]^2 + [1 + (\mathscr{k}/\alpha)^2][1 + (\alpha/k)^2]\sinh^2(\alpha L)}. \qquad (260)
\end{aligned}$$

If the limit $W \to 0$, then $\mathscr{k} \to k$ and the transmission coefficient reduces to that for a rectangular barrier:

$$\begin{aligned}
\mathscr{T} &= \frac{4}{4 + [1 + (k/\alpha)^2 + (\alpha/k)^2 + 1]\sinh^2(\alpha L)} \\
&= \frac{1}{1 + (1/4)[(k/\alpha) + (\alpha/k)]^2\sinh^2(\alpha L)}. \qquad (261)
\end{aligned}$$

This is the transmission coefficient for particles having energy $\mathscr{E} < U_0$ through a rectangular barrier of height U_0.

If the further limit $\alpha L \gg 1$, then $\sinh(\alpha L) \simeq \frac{1}{2}e^{\alpha L}$ and the approximate form is

$$\begin{aligned}
\mathscr{T} &\simeq \frac{1}{1 + (1/4)[(k/\alpha) + (\alpha/k)]^2(1/4)e^{2\alpha L}} \\
&\simeq \frac{16e^{-2\alpha L}}{[(k/\alpha) + (\alpha/k)]^2} = \left[\frac{4(\alpha/k)}{1 + (\alpha/k)^2}\right]^2 e^{-2\alpha L}. \qquad (262)
\end{aligned}$$

If the barrier thickness L approaches infinity, then the potential energy becomes a step potential and the transmission coefficient can be noted from Eq. (262) to approach zero for the presently considered case of $\mathscr{E} < U_0$.

To recapitulate, the remarkable quantum mechanical result has been derived that particles can, in some cases, penetrate potential energy barriers that are even higher than the particle energy. This result, which contrasts markedly with the classical result that $\mathscr{T} = 0$, has the greatest relevance for quantum electronic devices. It likewise provides the explanation for the decay of radioactive nuclei by α-particle emission.

An example calculation spanning the domains for $\mathscr{E} < U_0$ and $\mathscr{E} > U_0$, with $W = 0$ for both cases, has been carried out. The value of the electronic mass is utilized, and the rectangular barrier is chosen to have a thickness of 10 Å and a height of 10 eV. Figure 5 illustrates the variation of the transmission coefficient with incident electron energy. The remarkable oscillatory behavior is due to the wavelike nature of the particle; the peaks coincide with certain relationships between the de Broglie wavelength

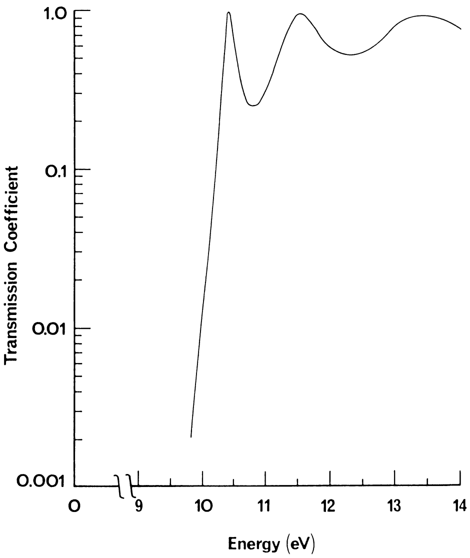


FIGURE 5 Transmission coefficient versus incident-particle energy for a rectangular potential energy barrier. [Fig. 1.35 in *Quantum Mechanics for Applied Physics and Engineering* by Albert Thomas Fromhold, Jr. (Academic Press, Inc., New York, 1981; Dover Publications, Inc., New York, 1991); reproduced with the permission of Academic Press, Dover Publications, and the author.]

and the barrier thickness, which can be deduced from Eq. (230). Note that $\mathcal{T}$ approaches unity whenever $\cos(2\beta L) = 1$, which occurs when $2\beta L = 2m\pi$ ($m = 1, 2, 3, \ldots$). Since $\beta = 2\pi/\lambda_{\text{II}}$, the condition is met when $m(\lambda_{\text{II}}/2) = L$ (i.e., whenever there are an integer number of half wavelengths over the barrier distance L). On the other hand, for $(2m+1)(\lambda_{\text{II}}/4) = L$, which corresponds to an odd-quarter wavelength over the barrier distance L, then $\cos(2\beta L) = -1$ and the transmission coefficient goes through the value $4/[(\beta/k) + (k/\beta)]^2$. In this odd-quarter wavelength condition, for $\beta \ll k$ corresponding to particle energies not far above the barrier maximum, $\mathcal{T} \simeq 4(\beta/k)^2$, which is very small in value. Thus, there are transmission resonances as the particle energy continuously increases above the barrier height U_0. For the odd-quarter wavelength condition with the particle energy far exceeding the barrier height, however, β is not too different from k and so $\mathcal{T}$ is not much less than unity. In short, the depth of the transmission resonances decreases as $\mathcal{E}$ increases.

It should be pointed out that although the odd-quarter wavelength condition usually provides a very good approximation for minimum $\mathcal{T}$, it does not provide the *exact* minimum. A quantitative evaluation based on $d\mathcal{T}/d\mathcal{E} = 0$ for fixed U_0, utilizing Eq. (230), leads to the condition

$$\tan(\beta L) = \frac{\beta L}{1 + (\beta/k)^2}$$

and the roots of this equation yield the distinct values of $\mathcal{E}$ for minimum $\mathcal{T}$. This transcendental equation can be solved graphically. The problem is simplified in the limit where $(\beta/k)^2 \ll 1$, since then the roots depend only on the energy difference above the barrier, thereby leading to pure wavelength conditions in the barrier region itself for the transmission minima. Further, in the large βL limit, the equation requires $\tan(\beta L)$ to be large, so that the roots approach values that are approximately the same as the values leading to the odd-quarter wavelength condition just discussed.

Finally, care must be taken not to draw the erroneous conclusion from a casual inspection of Eq. (230) that $\mathcal{T}$ is unity when $\beta = 0$, which corresponds to the particle energy being equal to the barrier height. A similar erroneous conclusion also could be reached from Eq. (261) in the same limit, since then α is zero. Careful inspection of these two equations reveals indeterminate forms in the denominators of the equations for this limit. In Eq. (230), for example, the two terms in the group $[(k/\beta)^2 - (k/\beta)^2 \cos(2\beta L)]$ approach $\infty \to \infty$ as $\beta \to 0$, but by noting the equivalence to $2(k/\beta)^2 \sin^2(\beta L)$, the indeterminate form yields $2(kL)^2$ as $\beta \to 0$. As given by Eq. (230), $\mathcal{T}$ therefore approaches $[1 + (kL/2)^2]^{-1}$ as $\beta \to 0$. In a similar manner, Eq. (261) can be shown to approach this same limit. The continuity in $\mathcal{T}$ at the limiting energy of 10 eV can be noted in Fig. 5.

E. Incident Beam with Particle Energy Below Both Steps

As a final consideration, let us examine the situation in which the particle energy $\mathcal{E}$ is less than the potential energy in region II and in region III. Algebraically, this is the case when $\mathcal{E} < U_0$ and $\mathcal{E} < W$. At the limit $W \to U_0$, this case reduces to the single step potential. In any event, since region III extends to $x = \infty$, the probability density in region III can be expected to approach zero as $x \to \infty$. The wavefunction for region III thus may be chosen to be

$$\psi_{\text{III}} = He^{-\gamma x}e^{-i\omega t}, \tag{263}$$

where

$$\gamma = \hbar^{-1}[2m(W - \mathcal{E}_0)]^{1/2}. \tag{264}$$

Matching the wavefunctions and the first derivatives at $x = 0$ yields the same results for the relationships among A, B, D, E as given by Eqs. (242)–(247), since the wavefunctions in regions I and II are exactly the same as for the case $W < \mathcal{E} < U_0$. However, the matching of the wavefunctions and their first derivatives at $x = L$ is different, since ψ_{III} is now different. The result obtained from this matching is

$$De^{-\alpha L} + Ee^{\alpha L} = He^{-\gamma L} \tag{265}$$

$$-\alpha De^{-\alpha L} + \alpha Ee^{\alpha L} = -\gamma He^{-\gamma L}. \tag{266}$$

Writing this pair of equations in the form

$$De^{-\alpha L} + Ee^{\alpha L} = He^{-\gamma L} \tag{267}$$

$$De^{-\alpha L} - Ee^{\alpha L} = \frac{\gamma}{\alpha}He^{-\gamma L} \tag{268}$$

facilitates the algebra. Adding these two equations leads to

$$D = \frac{1}{2}\left(1 + \frac{\gamma}{\alpha}\right)He^{-\gamma L} \tag{269}$$

and subtracting these two equations leads to

$$E = \frac{1}{2}\left(1 - \frac{\gamma}{\alpha}\right)He^{-\gamma L}. \tag{270}$$

Substituting these two expressions for D and E into the previously derived expressions for A and B—namely,

$$A = \frac{1}{2}\left[D\left(1 - \frac{\alpha}{ik}\right) + E\left(1 + \frac{\alpha}{ik}\right)\right] \tag{271}$$

$$B = \frac{1}{2}\left[D\left(1 + \frac{\alpha}{ik}\right) + E\left(1 - \frac{\alpha}{ik}\right)\right] \tag{272}$$

yields

$$
\begin{aligned}
A &= \frac{1}{2}\left\{\left(1-\frac{\alpha}{ik}\right)\left[\frac{1}{2}\left(1+\frac{\gamma}{\alpha}\right)He^{-\gamma L}\right]\right. \\
&\quad \left. +\left(1+\frac{\alpha}{ik}\right)\left[\frac{1}{2}\left(1-\frac{\gamma}{\alpha}\right)He^{-\gamma L}\right]\right\} \\
&= \frac{1}{4}\left[\left(1-\frac{\alpha}{ik}\right)\left(1+\frac{\gamma}{\alpha}\right)\right. \\
&\quad \left. +\left(1+\frac{\alpha}{ik}\right)\left(1-\frac{\gamma}{\alpha}\right)\right]He^{-\gamma L}. \qquad (273)
\end{aligned}
$$

$$
\begin{aligned}
B &= \frac{1}{2}\left\{\left(1+\frac{\alpha}{ik}\right)\left[\frac{1}{2}\left(1+\frac{\gamma}{\alpha}\right)He^{-\gamma L}\right]\right. \\
&\quad \left. +\left(1-\frac{\alpha}{ik}\right)\left[\frac{1}{2}\left(1-\frac{\gamma}{\alpha}\right)He^{-\gamma L}\right]\right\} \\
&= \frac{1}{4}\left[\left(1+\frac{\alpha}{ik}\right)\left(1+\frac{\gamma}{\alpha}\right)\right. \\
&\quad \left. +\left(1-\frac{\alpha}{ik}\right)\left(1-\frac{\gamma}{\alpha}\right)\right]He^{-\gamma L}. \qquad (274)
\end{aligned}
$$

The ratio of these two expressions gives

$$
\begin{aligned}
\frac{A}{B} &= \frac{[1-(\alpha/ik)][1+(\gamma/\alpha)]+[1+(\alpha/ik)][1-(\gamma/\alpha)]}{[1+(\alpha/ik)][1+(\gamma/\alpha)]+[1-(\alpha/ik)][1-(\gamma/\alpha)]} \\
&= \frac{[1+(\gamma/\alpha)+1-(\gamma/\alpha)]-(\alpha/ik)[1+(\gamma/\alpha)-1+(\gamma/\alpha)]}{[1+(\gamma/\alpha)+1-(\gamma/\alpha)]+(\alpha/ik)[1+(\gamma/\alpha)-1+(\gamma/\alpha)]} \\
&= \frac{2+i(\alpha/k)(2\gamma/\alpha)}{2-i(\alpha/k)(2\gamma/\alpha)} = \frac{[1+i(\gamma/k)]}{[1-i(\gamma/k)]}. \qquad (275)
\end{aligned}
$$

The reflection coefficient for this case is

$$
\mathcal{R} = \left(\frac{B}{A}\right)^{*}\left(\frac{B}{A}\right). \qquad (276)
$$

Substituting yields

$$
\begin{aligned}
\mathcal{R} &= \left[\frac{1-i(\gamma/k)}{1+i(\gamma/k)}\right]\left[\frac{1+i(\gamma/k)}{1-i(\gamma/k)}\right] \\
&= \frac{1+(\gamma^2/k^2)}{1+(\gamma^2/k^2)} = 1, \qquad (277)
\end{aligned}
$$

which is to be expected on physical grounds for this situation. These results have practical significance for modern quantum well devices based on multilayered semiconductors.

This completes the consideration of particle currents incident on stepfunction-type barriers. The basic approach also has some relevance for the simplest bound-state problem—namely, the problem of a particle trapped in a rectangular potential well—which is considered in Section VIII.

VIII. BOUND-STATE PROBLEMS

A. Introduction

A particle may be confined to a certain region of space by potential energy barriers surrounding it; e.g., the step potentials treated in Section VII could be placed on both sides of a particle having a lesser energy. Figure 6 illustrates an *arbitrary* potential energy function in one dimension. At positions x_1 and x_2, the potential energy $U(x)$ equals the total energy $\mathscr{E}$ of the particle, so that the kinetic energy $\mathscr{E}_k$, at these points given by $[\mathscr{E} - U(x)]$, is necessarily zero. These positions are called *classical turning points*, because according to classical physics, the particle would simply be reflected from the barrier at these positions. The momentum becomes reversed upon reflection. In three dimensions, the situation is similar, except that it is the perpendicular component of the momentum that is reversed (comparable to the elastic rebound of a rubber ball from a concrete wall). The motion of the particle continues back and forth over the region $x_1 < x < x_2$ in the potential well delineated by the surrounding energy barries, provided the potential energy everywhere outside this region exceeds the total particle energy. The total energy $\mathscr{E}$ is conserved, with a continuous interchange of kinetic and potential energies.

The detailed time-dependence of the motion of a particle in a potential energy well depends upon the exact functional form of the potential energy $U(x)$. The characteristic

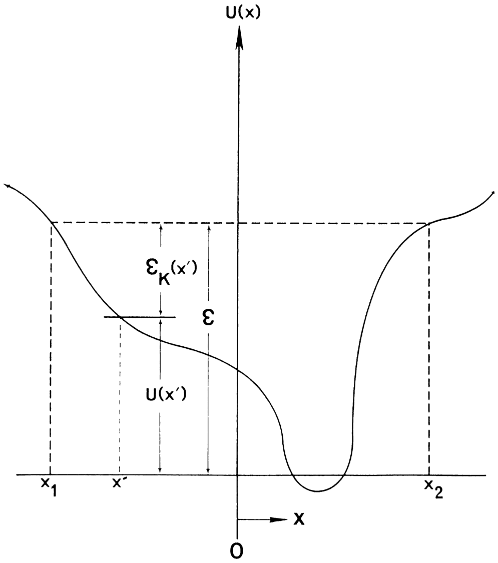


FIGURE 6 Potential energy well of arbitrary shape. [Fig. 1.38 in *Quantum Mechanics for Applied Physics and Engineering* by Albert Thomas Fromhold, Jr. (Academic Press, Inc., New York, 1981; Dover Publications, Inc., New York, 1991); reproduced with the permission of Academic Press, Dover Publications, and the author.]

feature of bound-state problems in quantum mechanics is the fact that the application of realistic boundary conditions to the time-independent Schrödinger equation, given by Eq. (169), forces a restriction on the energy values, so that the eigenvalues for the total energy are confined to a discrete set. This feature is independent of the particular functional form of the potential energy, although the details of the energy-level spectrum are dependent upon the form of $U(x)$. In the following discussion, three different bound-state problems, for which the potential is given by squarewell, harmonic oscillator, and Coulomb potentials, are considered individually.

B. Three-Dimensional Potential Energy Square-Well Problem

Let us consider a particle of mass m confined to a region of space having the shape of a rectangular parallelepiped. The particle confinement is due to infinite potential energy barriers at the faces of the parallelepiped box. This is sometimes referred to as a square-well potential because the potential energy rises so sharply (with infinite slope) at the boundaries of the parallelepiped. If the potential energy inside the parallelepiped is taken to be zero, then the Hamiltonian $\mathscr{H}$ for the time-independent and time-dependent Schrödinger equations [$\mathscr{H}\phi = \mathscr{E}\phi$ and $\mathscr{H}\psi = i\hbar\ \partial\psi/\partial t$, given by Eqs. (169) and (179)] takes the form $\mathscr{H} = -(\hbar^2/2m)\nabla^2$ within the box but is undefined outside the box, where the potential energy is considered infinite. Thus, a wavefunction ψ exists for the interior but is zero for the exterior, where the particle cannot penetrate. The time-independent Schrödinger equation for the interior of the box is formally the same as that for a free particle, namely,

$$-\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right)\phi_{\mathbf{k}}(\mathbf{r}) = \mathscr{E}_{\mathbf{k}}\phi_{\mathbf{k}}(\mathbf{r}), \tag{278}$$

which has solutions given by the normalized plane-wave spatial functions

$$\phi_{\mathbf{k}}(\mathbf{r}) = \left(\frac{1}{V}\right)^{1/2} \exp[i\mathbf{k}\cdot\mathbf{r}], \tag{279}$$

where V is the volume of the box. These spatial functions, when combined with the time factor $\exp[-(i/\hbar)\mathscr{E}_{\mathbf{k}}t]$, represent running waves

$$\psi_{\mathbf{k}}(\mathbf{r}, t) = \phi_{\mathbf{k}}(\mathbf{r}) \exp[(-i/h)\mathscr{E}_{\mathbf{k}}t], \tag{280}$$

which constitute the stationary-state wavefunctions for the particle that satisfy the time-dependent Schrödinger equation. These traveling-wave eigenfunctions of the Hamiltonian having the form $\exp[(i/\hbar)(\mathbf{p}\cdot\mathbf{r} - \mathscr{E}t)]$ are simultaneous eigenfunctions of the linear momentum operator $p^{op} = -i\hbar\nabla$, since it can be noted by direct substitution that

$$p^{op}\psi_{\mathbf{k}}(\mathbf{r}, t) = (-i\hbar)[\nabla(i\mathbf{k}\cdot\mathbf{r})]\psi_{\mathbf{k}}(\mathbf{r}) = \hbar\mathbf{k}\psi_{\mathbf{k}}(\mathbf{r}). \tag{281}$$

Even though such plane-wave solutions formally do satisfy the Schrödinger equation for this problem, they do not satisfy the constraint that the probability density be continuous at the walls of the box, since continuity of the wavefunction requires the wavefunction to be zero inside the box at the walls in order to equal the zero wavefunction exterior to the box. This underscores the vital role of boundary conditions in quantum mechanical solutions. As described in Section IV.F, however, oppositely directed but equal-amplitude plane waves can be superimposed to give standing-wave solutions that may satisfy the desired boundary conditions. Let us consider the six infinitely high potential energy barriers delimiting the rectangular parallelepiped box to be perpendicular to the x, y, and z axes and to be located at $x = 0, x = L_x, y = 0, y = L_y, z = 0, z = L_z$. The box confining the particle thus has edges of length L_x, L_y, and L_z in the x, y, and z directions, respectively. One form of the three-dimensional, normalized standing waves providing solutions to the Schrödinger equation for this problem is given by

$$\phi_{\mathbf{n}}\mathbf{r} = \left(\frac{8}{V}\right)^{1/2} \sin\left(\frac{n_x\pi x}{L_x}\right)\sin\left(\frac{n_y\pi y}{L_y}\right)\sin\left(\frac{n_z\pi z}{L_z}\right), \tag{282}$$

where n_x, n_y, and n_z are a triplet of positive integers represented by the symbol $\mathbf{n}$. This product of sine functions satisfies the fixed boundary condition that the wavefunction vanish on six faces of a rectangular parallelepiped, and direct substitution shows that it satisfies the three-dimensional Schrödinger equation given by Eq. (278). The corresponding quantized energy eigenvalues $\mathscr{E}_{\mathbf{n}}$ thereby obtained are

$$\mathscr{E}_{\mathbf{n}} = \left(\frac{\hbar^2\pi^2}{2m}\right)\left[\left(\frac{n_x}{L_x}\right)^2 + \left(\frac{n_y}{L_y}\right)^2 + \left(\frac{n_z}{L_z}\right)^2\right]. \tag{283}$$

Only the positive integers are chosen for the triplet n_x, n_y, n_z in the wavefunction given by Eq. (282), since the corresponding negative values yield the same wavefunction to within a factor of -1. This would represent exactly the same state for all practical quantum mechanical calculation purposes, because the particle probability density $\psi^*\psi$ would be the same. It is a very general aspect of quantum mechanics that linearly dependent eigenfunctions are redundant whereas linearly independent eigenfunctions are not.

The standing-wave solutions of the Schrödinger equation given by Eq. (282) do not represent states having definite momentum values, but instead are linear combinations of plane-wave states having oppositely directed momenta. Because the particle undergoes reversals in momentum associated with reflection in the neighborhood of the classical turning points, it is not surprising that the most appropriate solutions to the problem are not those states that represent definite momentum values.

Because the standing-wave solutions can be viewed as the superposition of traveling-wave solutions of equal amplitude traveling in opposite directions, plane-wave solutions can be utilized for the present problem by means of a construct known as periodic boundary conditions:

$$\begin{aligned}\phi(x+L_x, y, z) &= \phi(x, y, z)\\ \phi(x, y+L_y, z) &= \phi(x, y, z)\\ \phi(x, y, z+L_z) &= \phi(x, y, z).\end{aligned} \tag{284}$$

This represents a substitute for the fixed boundary conditions previously invoked to determine the allowable **k** values. This is the approach generally used in the development of the free-electron theory of metals. It has the utility of simplifying the mathematics a bit while retaining the essential property of the correct number of quantum states per unit energy range required for quantum statistical calculations.

C. The Harmonic Oscillator Potential

Another very important bound-state potential energy function is the harmonic oscillator potential

$$U(x) = \frac{1}{2}Kx^2 \tag{285}$$

illustrated in Fig. 7. For a one-dimensional harmonic oscillator consisting of a mass M attached to a fixed spring (with force constant K), which is set into motion on a horizontal; frictionless planar surface, the classical frequency given by the reciprocal of the period is $\nu_0 = \omega_0/2\pi = (2\pi)^{-1}(K/M)^{1/2}$. Other examples of classical harmonic oscillators also are characterized by some specific frequency (with corresponding angular frequency $\omega_0 = 2\pi\nu_0$) determined by the physical parameters of the system in question. The analogous quantum problem has solutions that can be expressed in terms of this classical frequency ν_0.

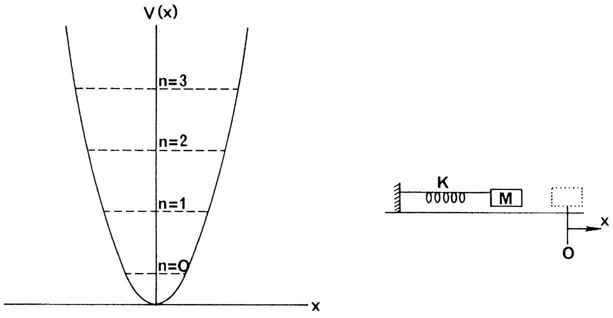


FIGURE 7 Harmonic oscillator potential. [Fig. 1.39 in *Quantum Mechanics for Applied Physics and Engineering* by Albert Thomas Fromhold, Jr. (Academic Press, Inc., New York, 1981; Dover Publications, Inc., New York, 1991); reproduced with the permission of Academic Press, Dover Publications, and the author.]

The Hamiltonian for the harmonic oscillator problem is

$$\mathscr{H} = \left(-\frac{\hbar^2}{2m}\right)\left(\frac{d^2}{dx^2}\right) + \frac{1}{2}Kx^2 \tag{286}$$

and the time-independent Schrödinger equation given by Eq. (169) for this problem is

$$\left(-\frac{\hbar^2}{2m}\right)\left[\frac{d^2\phi_n(x)}{dx^2}\right] + \frac{1}{2}Kx^2\phi_n(x) = \mathscr{E}_n\phi_n(x). \tag{287}$$

The complete time-dependent solutions $\psi_n(x, t)$ then are given as usual, by taking the product of the spatial functions with the corresponding time-dependent function $\theta_n(t) = \exp[(-i/\hbar)\mathscr{E}_n t]$, so that

$$\psi_n(x, t) = \phi_n(x)\exp\left[-\left(\frac{i}{\hbar}\right)\mathscr{E}_n t\right]. \tag{288}$$

In the present case

$$\phi_n(x) = \mathscr{N}_n e^{-\alpha x^2/2} H_n(\alpha^{1/2}x) \tag{289}$$

with

$$\alpha \equiv \frac{m\omega_0}{\hbar}. \tag{290}$$

The functions $H_n(y)$ appearing in Eq. (289), with $y \equiv \alpha^{1/2}x$, are the Hermite polynomials. These polynomials are readily generated by means of the differential relation

$$H_n(y) = (-1)^n \exp(y^2)\frac{d^n}{dy^n}[\exp(-y^2)]. \tag{291}$$

Some examples of the lower-order Hermite polynomials are

$$\begin{aligned}H_0(y) &= 1\\ H_1(y) &= 2y\\ H_2(y) &= 4y^2 - 2\\ H_3(y) &= 8y^3 - 12y\\ H_4(y) &= 16y^4 - 48y^2 + 12\\ H_5(y) &= 32y^5 - 160y^3 + 120y\\ H_6(y) &= 64y^6 - 480y^4 + 720y^2 - 120\\ H_7(y) &= 128y^7 - 1344y^5 + 3360y^3 - 1680y.\end{aligned} \tag{292}$$

The normalization factors $\mathcal{N}_n$ for the eigenfunctions, as found from the normalization integral

$$\int_{-\infty}^{\infty} \psi_n^* \psi_n \, dx = 1 \tag{293}$$

are given by

$$\mathcal{N}_n = \left[\frac{\alpha}{2^n n! \pi^{1/2}} \right]^{1/2}. \tag{294}$$

The various eigenfunctions $\phi_n(x)$ are orthogonal, this being a common occurrence in all quantum solutions representing different energies. Plots that illustrate these lower-order harmonic oscillator eigenfunctions can be found in various textbooks. The average amplitude for the vibrational motion increases with increasing-energy eigenvalues, in qualitative agreement with the predictions of the classical treatment. Beyond this point, there is little apparent similarity unless one examines the solutions in the so-called correspondence limit of very large quantum numbers.

The requirement that acceptable wavefunctions ordinarily must be capable of normalization is what actually leads to the Hermite polynomial solutions. This requirement simultaneously leads to the discrete spectrum of quantized energy eigenvalues

$$\begin{aligned} \mathcal{E}_n &= \left(n + \frac{1}{2}\right)\hbar\omega_0 \\ &= \left(n + \frac{1}{2}\right)h\nu_0 \qquad (n = 0, 1, 2, \ldots, \infty), \end{aligned} \tag{295}$$

where ω_0 is the angular frequency for the classical mechanics solution. The integer n is the quantum number for this problem. An interesting feature of the quantum solution is the fact that the ground-state energy ($n = 0$) is nonzero. Furthermore, the energy levels are evenly spaced. The dashed lines in Fig. 7 indicate this energy-level spectrum.

The topic of lattice vibrations in solids is one example of an important physical problem that can be treated in terms of the harmonic oscillator. The fact that the ground-state energy is nonzero means that even at a temperature of zero Kelvin, there will be some vibrational motion of the lattice. Such motions are the so-called zero-point vibrations.

Another interesting feature of the quantum solution is the dependence of the energy upon the frequency of oscillation. The increase in amplitude (see Fig. 7) that accompanies larger values of the energy $\mathcal{E}_n$ for the classical solution seems almost incidental to the quantum solution, whereas in the classical solution, the dependence of energy upon amplitude appears as a central feature in the analysis.

The energy $\Delta\mathcal{E}(n \to n')$ required to excite a quantum oscillator of frequency ω from a state of quantum number n to the state of quantum number n' can be noted from Eq. (295) to be

$$\begin{aligned} \Delta\mathcal{E}(n \to n') &= \left(n' + \frac{1}{2}\right)\hbar\omega - \left(n + \frac{1}{2}\right)\hbar\omega \\ &= (n' - n)\hbar\omega. \end{aligned} \tag{296}$$

Thus, energy absorption occurs in integer multiples of a basic unit of energy $\hbar\omega$, which is characteristic of the oscillator frequency. Analogously, the energy emission due to the deexcitation of an oscillator of frequency ω from a state of quantum number n' to the state of quantum number n is the converse of the absorption process. If the energy is emitted in the form of a photon of energy $\mathcal{E}_{\text{photon}} = h\nu_{\text{photon}}$, then the frequency of the photon will be $\nu_{\text{photon}} = (n' - n)\,\omega/2\pi = (n' - n)\nu_{\text{osc}}$ and the wavelength of the emitted photon will be $\lambda_{\text{photon}} = c/\nu_{\text{photon}} = c/[(n' - n)\nu_{\text{osc}}]$. This leads to the quantum transition picture indicated in Fig. 2.

D. Use of the Schrödinger Equation for the Hydrogen Atom and One-Electron Ions

The first difficulty encountered in the use of classical mechanics was in the area of radiation absorption and emission by atoms in gases. Instead of the continuous spectra predicted by the classical mechanics approach (see Section I.D), discrete optical spectra were obtained (as described in Section III.D). Although the 1913 semiclassical approach of Bohr (see Section III.F) was successful in explaining such discrete spectra for hydrogen and also for one-electron ions, there was no way to extend the Bohr theory to the two-electron atom (helium) or similar two-electron ions, much less to even higher electron atoms and ions. An approach that could explain only such a limited region of the periodic table (i.e., one element—namely, hydrogen) obviously was inadequate and ultimately had to be replaced by a more general theory. In 1926, Schrödinger developed what was to be the genesis of such a more general approach [see Section III.E].

Let us now devote some attention to the solution of the Schrödinger equation for the hydrogen atom and the closely related problem of the one-electron ion, the potential energy for both being given by

$$U(r) = \frac{-Ze^2}{4\pi\varepsilon_0 r}, \tag{297}$$

where Ze is the electrical charge of the nucleus and $r = |\mathbf{r}|$ is the separation distance between the centers of the two charges. This potential energy as a function of separation distance appears as solid curves in Fig. 8. Although this is formally a two-body problem, it can be reduced to a one-body problem by transforming to the center-of-mass coordinate system. The so-called "reduced" mass of the

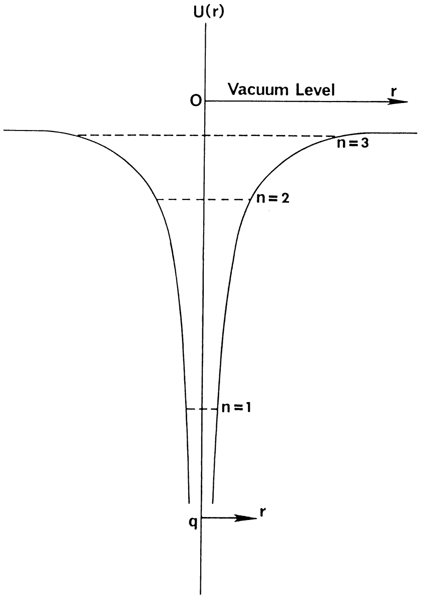


FIGURE 8 Coulomb potential. [Fig. 1.40 in *Quantum Mechanics for Applied Physics and Engineering* by Albert Thomas Fromhold, Jr. (Academic Press, Inc., New York, 1981; Dover Publications, Inc., New York, 1991); reproduced with the permission of Academic Press, Dover Publications, and the author.]

electron given by

$$m = \frac{m_e m_p}{m_e + m_p}, \tag{298}$$

where m_e is the actual electron rest mass and m_p is the proton rest mass. The Hamiltonian

$$\mathscr{H} = -\frac{\hbar^2}{2m}\nabla^2 + \frac{q_1 q_2}{4\pi\varepsilon_0 r} \tag{299}$$

with nuclear charge $q_1 = Z_1 e$ and electron charge $q_2 = -e$, is to be used in the time-independent Schrödinger equation $\mathscr{H}\phi_n = \mathscr{E}_n\phi_n$. The key to solving this so-called hydrogen-atom problem is the recognition that the Coulomb potential energy of interaction between the electron in question and the nucleus depends only upon the separation distance $r \equiv |\mathbf{r}|$ and is independent of the spatial orientation of the line of centers between electron and nucleus. Therefore, in spherical polar coordinates (r, θ, ϕ), the Schrödinger equation can be separated (in the usual way variables are separated in partial differential equations) into three equations, each involving a function of one of these three variables. In spherical polar coordinates, the Laplacian ∇^2 of any arbitrary scalar function $\mathscr{f}$ of the vector $\mathbf{r}$ is given by

$$\nabla^2\mathscr{f} = \left(\frac{1}{r^2}\right)\left(\frac{\partial}{\partial r}\right)\left(r^2\frac{\partial\mathscr{f}}{\partial r}\right) + \left(\frac{1}{r^2\sin\theta}\right)\left(\frac{\partial}{\partial\theta}\right) \times \left[(\sin\theta)\frac{\partial\mathscr{f}}{\partial\theta}\right] + \left(\frac{1}{r^2\sin^2\theta}\right)\left(\frac{\partial^2\mathscr{f}}{\partial\phi^2}\right). \tag{300}$$

The variables separation technique then leads to three separated equations for the factors $R(r)$, $\Theta(\theta)$, and $\Lambda(\phi)$ appearing in the product form of the spatial portion of the wavefunction $\phi(\mathbf{r}) = R(r)\Theta(\theta)\Lambda(\phi)$, with the corresponding wavefunction being

$$\psi(\mathbf{r}, t) = R(r)\Theta(\theta)\Lambda(\phi)\exp\left(\frac{-i}{\hbar}\mathscr{E}t\right). \tag{301}$$

Only the equation for $R(r)$ contains the Coulomb potential energy of interaction between electron and nucleus.

Single-valuedness of the wavefunction is necessary because the probability density $\psi^*\psi$ is a physical quantity that must have only one value at a given point in space. The separation constant in the equation for $\Lambda(\phi)$ is found to lead to a wavefunction $\psi(\mathbf{r}, t) \propto \exp(im\phi)$, which is single-valued whenever the angle ϕ is increased by multiples of 2π only if m is equal to an integer. The m constitutes one quantum number characterizing the electronic state. It is called the magnetic quantum number, since its value determines the energy change when the ion is placed in a magnetic field. It is conceivable that m has allowable values of $0, \pm1, \pm2, \ldots,$ although an upper bound on $|m|$ is dictated by another consideration (to be discussed shortly). The resulting wavefunction factor $\Lambda(\phi)$ is found to be an eigenfunction of the Hermitian operator $\mathscr{L}_{\mathfrak{z}}$, which represents the $\mathbf{z}$-component of the electron orbital angular momentum, with the eigenvalue being $m\hbar$:

$$\mathscr{L}_{\mathfrak{z}}\Lambda\phi = m\hbar\,\Lambda(\phi) \qquad (m = 0, \pm1, \pm2, \ldots). \tag{302}$$

The differential equation for $\Theta(\theta)$ representing the θ-component of the wavefunction $\psi(\mathbf{r}, t)$ contains m^2 as well as a second separation constant λ. Whenever $m = 0$, the equation can be cast into a form known as Legendre's differential equation. The solutions diverge unless $\lambda = \ell(\ell + 1)$, where ℓ represents a non-negative integer. The physically meaningful probability densities obtained for this choice of the second separation constant are represented by the solutions $P_\ell(\cos\theta)$, known as the Legendre polynomials. These polynomials turn out to be eigenfunctions of the Hermitian operator $[\mathscr{L}^{(\mathrm{op})}]^2$representing the square of the orbital angular momentum of the electron, with eigenvalues $\ell(\ell + 1)\hbar^2$:

$$\left[\mathscr{L}^{(\mathrm{op})}\right]^2\Theta(\theta) = \ell(\ell + 1)\hbar^2\Theta(\theta) \qquad (\ell = 0, 1, 2, 3, \ldots). \tag{303}$$

Therefore, ℓ is called the orbital angular momentum quantum number.

Whenever $m \neq 0$, the corresponding solutions to the $\Theta(\theta)$ equation are the associated Legendre functions $P_\ell^m(\cos\theta)$, where $m^2 \leq \ell^2$. The restriction on m is thus a mathematical one, although it ties in very well with the corresponding physics since the square of the **z**-component of the orbital angular momentum—namely, $m^2\hbar^2$—cannot exceed the square of the total orbital angular momentum—namely, $\ell(\ell+1)\hbar^2$—which, in turn requires $|m| \le \ell$.

The differential equation for $R(r)$ representing the r-component of the wavefunction $\psi(\mathbf{r}, t)$ contains the quantum number ℓ as well as a third separation constant. The solutions can be expressed in terms of the associated Laguerre polynomials. The requirement that the solutions not diverge in order to have a physically meaningful probability density $\psi^*\psi$ once again places severe restrictions on the separation constant. This, in turn, requires an integer quantum number n, known as the principal quantum number, together with the condition $n > \ell$. In a straightforward way, this leads to the quantized energy eigenvalues

$$\mathscr{E}_n = -\frac{1}{2}\frac{Z^2e^2}{4\pi\varepsilon_0 a_0 n^2} \qquad (n = 1, 2, 3, \ldots), \tag{304}$$

where a_0 is the parameter known as the *Bohr radius*

$$a_0 = \frac{4\pi\varepsilon_0\hbar^2}{me^2} \tag{305}$$

since it equals the radius of the ground-state Bohr circular orbit previously deduced from Eq. (31) for $n = 1$. The energy levels for the lower-energy states are indicated by dashed lines in Fig. 8.

From a different perspective, the hydrogenic eigenfunctions resulting from the mathematical solution to this problem can be written in terms of the product

$$\begin{aligned} \psi_{n\ell_m} &= R_n\ell(r)Y\ell_m(\theta, \phi) \\ &(n = 1, 2, 3, \ldots; \quad 0 \le \ell \le n-1; \\ &m = 0, \pm1, \pm2, \ldots, \pm\ell), \end{aligned} \tag{306}$$

where $R_n\ell(r)$is the radial portion of the eigenfunction (which is correlated directly with the functional form of the Coulomb potential energy) and $Y\ell_m(\theta, \phi) = \Theta\ell_m(\theta)\Lambda_m(\phi)$ is one of the spherical harmonics. The normalized spherical harmonics $Y\ell_m$ can be written in terms of the associated Legendre polynomials $P_\ell^m(\cos\theta)$:

$$Y\ell_m(\theta, \phi) = \left[\frac{(2\ell+1)}{4\pi}\frac{(\ell-m)!}{(\ell+m)!}\right]^{1/2} P_\ell^m(\cos\theta)e^{im\phi}. \tag{307}$$

Several of the lower-order associated Legendre polynomials are

$$\begin{aligned}
\ell = 0 \qquad & P_0^0(\cos\theta) = 1 \\
\ell = 1 \qquad & P_1^0(\cos\theta) = \cos\theta \\
& P_1^1(\cos\theta) = \sin\theta \\
\ell = 2 \qquad & P_2^0(\cos\theta) = \tfrac{1}{2}(3\cos^2\theta - 1) \\
& P_2^1(\cos\theta) = 3\sin\theta\cos\theta \\
& P_2^2(\cos\theta) = 3\sin^2\theta \\
\ell = 3 \qquad & P_3^0(\cos\theta) = \tfrac{1}{2}(\cos\theta)(5\cos^2\theta - 3) \\
& P_3^1(\cos\theta) = \tfrac{3}{2}(\sin\theta)(5\cos^2\theta - 1) \\
& P_3^2(\cos\theta) = 15\sin^2\theta\cos\theta \\
& P_3^3(\cos\theta) = 15\sin^3\theta \\
\ell = 4 \qquad & P_4^0(\cos\theta) = \tfrac{1}{8}(35\cos^4\theta - 30\cos^2\theta + 3) \\
& P_4^1(\cos\theta) = \tfrac{5}{2}(\sin\theta)(7\cos^3\theta - 3\cos\theta) \\
& P_4^2(\cos\theta) = \tfrac{15}{2}(\sin^2\theta)(7\cos^2\theta - 1) \\
& P_4^3(\cos\theta) = 105\sin^3\theta\cos\theta \\
& P_4^4(\cos\theta) = 105\sin^4\theta.
\end{aligned} \tag{308}$$

To obtain the corresponding polynomials for the negative m values, the relation

$$P_\ell^{-m}(\cos\theta) = (-1)^m\frac{(\ell-m)!}{(\ell+m)!}P_\ell^m(\cos\theta) \tag{309}$$

can be utilized. The complete set of these polynomials can be generated in a straightforward manner.

The spherical harmonics actually are appropriate for all spherically symmetric potential energy problems. For this reason, they are used as the angular portion of the wavefunction in many approximate treatments of the many-electron atom. Furthermore, the use of spherical harmonics is by no means restricted to quantum mechanics; e.g., they are utilized extensively in treating boundary-value problems in electrostatics. The usefulness derives from the fact that they constitute a *complete* set of functions, so that any arbitrary function $g(\theta, \phi)$ of θ and ϕ can be expanded as a linear combination of the spherical harmonics:

$$g(\theta, \phi) = \sum_{\ell=0}^{\infty}\sum_{m=-\ell}^{\ell} a_{\ell m}Y_{\ell m}(\theta, \phi). \tag{310}$$

The coefficients $a_{\ell m}$ in the linear combination are determined from

$$a_{\ell m} = \iint Y_{\ell m}^*(\theta, \phi)g(\theta, \phi)\,d\mathscr{S} \tag{311}$$

with the differential $d\mathscr{S}$ denoting the differential surface area on a unit sphere $d\mathscr{S} = \sin\theta\,d\theta\,d\phi$, with the limits of integration being 0 to π for θ and 0 to 2π for ϕ.

The normalized radial portion of the energy eigenfunctions for hydrogenlike atoms can be written in the form

$$R_{n\ell}(r) = \left[\left(\frac{2Z}{na_0}\right)^3 \left(\frac{(n-\ell-1)!}{[2n(n+\ell)!]^3}\right)\right]^{1/2} e^{\rho/2}\rho^{\ell} L_{n+\ell}^{2\ell+1}(\rho), \tag{312}$$

where

$$\rho \equiv 2\left(\frac{-2m\mathscr{E}}{\hbar}\right)^{1/2} r. \tag{313}$$

The parameter a_0 is the Bohr radius given by Eq. (31) with $n = 1$, and $L_{n+\ell}^{2\ell+1}$ represents the associated Laguerre polynomials, which can be generated from the differential relation

$$L_{n+\ell}^{2\ell+1}(\rho) = \frac{d^{2\ell+1}}{d\rho^{2\ell+1}}\left[e^{\rho}\frac{d^{n+\ell}}{d\rho^{n+\ell}}(\rho^{n+\ell}e^{-\rho})\right] \tag{314}$$

The lower-order Laguerre polynomials are therfore easily obtained, thereby yielding the radial portions of the one-electron energy eigenfunctions. Several of the lower-order radial function are

$$\begin{aligned}
n = 1 \quad & R_{10}(r) = 2\left(\frac{Z}{a_0}\right)^{3/2} e^{-\rho/2} \\
n = 2 \quad & R_{20}(r) = (8)^{-1/2}\left(\frac{Z}{a_0}\right)^{3/2}(2-\rho)e^{-\rho/2} \\
& R_{21}(r) = (24)^{-1/2}\left(\frac{Z}{a_0}\right)^{3/2} e^{-\rho/2} \\
n = 3 \quad & R_{30}(r) = (243)^{-1/2}\left(\frac{Z}{a_0}\right)^{3/2}(6-6\rho+\rho^2)e^{-\rho/2} \\
& R_{31}(r) = (486)^{-1/2}\left(\frac{Z}{a_0}\right)^{3/2}(4-\rho)e^{-\rho/2} \\
& R_{32}(r) = (2430)^{-1/2}\left(\frac{Z}{a_0}\right)^{3/2}\rho^2 e^{-\rho/2} \qquad (315) \\
n = 4 \quad & R_{40}(r) = (96)^{-1}\left(\frac{Z}{a_0}\right)^{3/2} \\
& \qquad \times (24 - 36\rho + 12\rho^2 - \rho^3)e^{-\rho/2} \\
& R_{41}(r) = [15 \times (32)^2]^{-1/2}\left(\frac{Z}{a_0}\right)^{3/2} \\
& \qquad \times (20 - 10\rho + \rho^2)e^{-\rho/2} \\
& R_{42}(r) = [5 \times (96)^2]^{-1/2}\left(\frac{Z}{a_0}\right)^{3/2}(6-\rho)\rho^2 e^{-\rho/2} \\
& R_{43}(r) = [35 \times (96)^2]^{-1/2}\left(\frac{Z}{a_0}\right)^{3/2}\rho^3 e^{-\rho/2}.
\end{aligned}$$

To summarize the situation for the one-electron ion, the application of appropriate boundary conditions of single-valuedness and boundedness on the wavefunction leading to the preceding solutions yields the three quantum numbers n, ℓ, and m, which are referred to, respectively, as the principal quantum number, the orbital (or azimuthal) quantum number, and the magnetic quantum number. The allowed values of the three integer quantum numbers (n, ℓ, m) required to characterize a given energy eigenfunction $\psi(\mathbf{r}, t) = \psi_m\ell_m(\mathbf{r}, t)$ for the hydrogen atom or one-electron ion have the following allowed values:

$$\begin{aligned}
n &= 1, 2, 3, \ldots \\
0 &\leq \ell \leq n-1 \qquad (316) \\
-\ell &\leq m \leq \ell.
\end{aligned}$$

It can be seen from these results that for a given n-value, there are n allowable ℓ values and for a given ℓ-value, there are $2\ell + 1$ possible m values. Since the energy eigenvalues $\mathscr{E}_n$ given by Eq. (304) depend only on the value of the principal quantum number n and are independent of the values of ℓ and m, it can be seen that there are many eigenfunctions for a given energy eigenvalue. The solutions for the hydrogen atom are therefore highly degenerate, except for the ground state ($n = 1$, $\ell = 0$, $m = 0$). When the fourth quantum number m_s, representing electron spin, is taken into account ($m_s = \pm 1/2$), corresponding to spin angular momentum values of $\pm m_s\hbar$, two electrons can be accommodated in the ground state without violating the Pauli exclusion principle, developed by Wolfgang Pauli (1900–1958). That is, no two electrons have exactly the same set of quantum numbers.

Due to the degeneracy of the eigenfunctions, it is possible to construct new linear combinations for a given value of the principal quantum number n (i.e., for a given shell). Such alternative sets become extremely important where there are additional contributions to the energy that can be treated as perturbations. The appropriate choice of the basis set is often dictated by the symmetry of the perturbation.

The series of energy levels is modified in the presence of a magnetic field. There are two factors to consider—namely, the electron orbital angular momentum **L** for motion around the nucleus, and the intrinsic spin of the electron. A charged particle in orbit constitutes a circulating current, which, in turn, produces a magnetic moment. This so-called orbital magnetic moment $\boldsymbol{\mu}_{\text{orb}} = -(e/2m)\mathbf{L}$ interacts with an applied magnetic field **B** to give an additional energy term

$$U_{\text{orb}} = -\boldsymbol{\mu}_{\text{orb}} \cdot \mathbf{B}, \tag{317}$$

which must be added to the Hamiltonian. With the magnetic field $\mathbf{B} = B_z\hat{\mathbf{z}}$ oriented along the **z**-axis, the energy is $m\hbar B_z$, where the integer m is the magnetic quantum number. Likewise, the intrinsic spin angular momentum of the electron with respect to an axis through its center of mass gives rise to a spin magnetic moment $\boldsymbol{\mu}_s$, which interacts with an applied magnetic field **B** to give the energy term

$$U_{\text{spin}} = -\boldsymbol{\mu}_s \cdot \mathbf{B}, \tag{318}$$

which must be added to the Hamiltonian. Other energy terms also can arise, such as the energy of the interaction between the two magnetic moments $\boldsymbol{\mu}_{\text{orb}}$ and $\boldsymbol{\mu}_s$ that is designated the spin–orbit interaction energy.

E. Many-Electron Atoms and Ions

First of all, it is important to recognize that the potential energy in a multi-electron atom depends upon the electron–electron Coulomb energies as well as on the Coulomb energy of interaction of each electron with the nucleus. Although the electron–electron interaction can be viewed as a perturbation for purposes of very crude estimates, it, in fact, is much too large to be treated within the framework of perturbation theory. In the helium atom, for example, the potential energy involves the Coulomb potential energy of interaction of the attractive force between each electron and the nucleus plus the Coulomb energy of interaction of the repulsive force between the two electrons. The wavefunction ψ then must be taken at minimum as the product of the wavefunctions for each individual electron in the atom. The problem becomes enormously more complicated to solve, requiring a numerical solution because an exact analytical solution cannot be found. Nevertheless, the Schrödinger equation yields numerical results for the two-electron atom problem that agree with experimental optical spectra. In this way, the Schrödinger equation provides the more general theory required to supplant the more limited Bohr theory.

The complexity of an exact numerical solution of the Schrödinger equation naturally increases greatly as the number of electrons increases from two three. If only a few electrons are involved, as in the case for the lighter atoms, a variational treatment often can be used to obtain approximate solutions to the many-electron Schrödinger equation.

For the heavier atoms, wherein a larger number of elecrons are involved, the starting point for the most calculations is the approximation that the total potential energy of interaction of a given electron with the nucleus and the other electrons can be represented by a spherically symmetric potential $U(\mathbf{r})$, referred to as the central-field approximation. In practice, then, one of the most difficult parts of the problem is to estimate or calculate the potential. The details are beyond the scope of the present treatment; however, the results justify utilizing one-electron eigenvalues and eigenfunctions as a semantic framework for describing the multi-electron atom or ion. Thus, the product wavefunction

$$\psi(\mathbf{r}, t) = R(r)\Theta(\theta)\Lambda(\phi)\exp\left(\frac{-i}{\hbar}\mathscr{E}t\right) \tag{319}$$

is chosen to have the same form as that given by Eq. (301) for the one-electron atom. Once again, a set of four quantum numbers (n, ℓ, m_ℓ, and m_s) is required to specify an electronic state. The wavefunction specified by a given set of quantum numbers is called an orbital or, more specifically, an atomic orbital, in analogy with the older Bohr theory in which electrons were considered to travel in planetary orbits in accordance with classical mechanics.

The Pauli exclusion principle requires that no two electrons have the same set of quantum numbers [viz., the principal quantum number n characteristic of the total energy of the electron, the angular momentum (azimuthal) quantum number ℓ characteristic of the total orbital angular momentum of the electron, the magnetic quantum number m_ℓ characteristic of the orientation of the magnetic moment with respect to the **z**-azis, and the spin quantum number m_s characteristic of the orientation of the electron-spin magnetic moment]. The orbital angular momentum and magnetic quantum numbers ℓ and m_ℓ for the multi-electron atom or ion are the same as the quantum numbers ℓ and m in the hydrogen atom, since the variables separation with the more general central potential $U(r)$ proceeds in exactly the same way as for the one-electron atom, for which $U(r) = -Ze^2/4\pi\varepsilon_0 r$, thereby yielding the same equations for the $\Theta(\theta)$ and $\Lambda(\phi)$ factors in $\psi(\mathbf{r}, t)$. The electron-spin quantum number $m_s = \pm 1/2$ is likewise the same as in the hydrogen atom. The radial equation, containing as it does the generalized central potential $U(r)$ instead of simply the electron–nucleus Coulomb potential, requires a generalized *total* quantum number n analogous to the *principal* quantum number n for the hydrogen atom.

One very important difference between the results for the general central potential problem, in which the potential no longer varies as $1/r$, and the hydrogen atom problem, in which the potential varies strictly as $1/r$, is the fact that electronic states characterized by different values of the orbital angular momentum quantum number ℓ with the same total quantum number n generally correspond to different energy eigenvalues, whereas in the hydrogen-atom problem, the energy eigenfunctions for a given n but different ℓ values are degenerate. In multi-electron atoms, states of lower ℓ-value consistent with a fixed n-value lie at a lower energy. The combined values of ℓ and n for a given eigenfunction determine the radial nodes, these being $n - \ell - 1$ in number. As in the hydrogen atom, n must be a positive integer, and the magnitude of the integer ℓ cannot exceed $n - 1$. An atomic shell is specified by a given value for n, and an atomic subshell is specified by a given set of values for both n and ℓ. Taking into account the two possible spin quantum numbers $m_s = \pm 1/2$ and the $(2\ell + 1)$ values for $m_\ell[m_\ell = -\ell, -\ell + 1, \ldots 0, 1, \ldots \ell]$, one deduces the result that a given subshell contains $2(2\ell + 1)$ degenerate electronic states. In standard spectroscopic notation, the series of shells are denoted by $K, L, M, N, \ldots.$

The ground state of a many-electron atom is the one in which a sufficient number of electrons populates the orbitals of lowest energy consistent with the Pauli exclusion principle to give a neutral entity. The ground-state configuration of the electrons in an atom is specified by the number of electrons in each shell. The chemical properties of the different atoms (or elements) are determined principally by the uppermost filled energy levels, since these higher-energy electrons, being less tightly bound to the atomic core, most easily share themselves with adjacent atomic cores for the formation of chemical bonds in molecules and solids. If the uppermost occupied shell is full, there is generally an appreciable difference in energy between the occupied and next-higher unoccupied state, and the atom then tends to be chemically inert.

It is standard in spectroscopic notation to give the n-value of a shall as a number and the ℓ-value as a lowercase letter, with $\ell = 0, 1, 2, 3, 4, \ldots$ being denoted, respectively, by the letters $s, p, d, f, g, \ldots$. The periodic filling of successive shells as Z increases explains the use of a periodic table for listing the chemical elements. The number of electrons in a given shell generally is denoted by a superscript. For example, sodium has two electrons in the 1s shell, two electrons in the 2s shell, six electrons in the 2p shell, and one electron in the 3s shell; this ground-state configuration for sodium ($Z = 11$) would be denoted by Na: $1s^2 2s^2 2p^6 3s$. The rule that the maximum number of electrons in a shell be $2(2\ell + 1)$, with $\ell \leq n - 1$, can be consulted in conjunction with this configuration to illustrate that atomic sodium consists of two filled shells (or three filled subshells) containing the core electrons and an outermost partly filled shell containing the single valence electron.

The use of one-electron states to characterize the electronic states of the multi-electron atom is a good illustration of how physical models for complicated systems are constructed. In this section, a simple framework, supplemented by the additional requirements of the problem in question, has enabled an understanding of a much more complex problem to be developed. In the present case, the simplifying assumption is that of a spherically symmetric potential and the additional condition is that of the Pauli exclusion principle. The overall approach permits a rudimentary understanding of the entire periodic table for the chemical elements.

Section IX examines the predictions of the Schrödinger equation for an electron in the presence of a periodic potential energy. The periodic potential represents another example of a simple framework used as the basis for models of a very complicated physical system. This case relies on the fact that periodicity in the potential is a common attribute of the atom array in a crystalline solid. The problem of electrons in a solid involves the population of many energy levels at one time, so the quantum statistics governing the behavior of many electrons in the same system again plays an important role. The treatment of the periodic-potential problem leads to energy band theory—an interesting and important topic, since it is so successful in explaining the difference between metals, semiconductors, and insulators.

IX. ELECTRON TRANSPORT IN SOLIDS

A. Failure of Classical Physics for Electrical Currents in Solids

Another important difficulty encountered in classical mechanics is in the area of electron transport in solids. Viewing condensed matter as merely an agglomeration of hard-sphere atoms packed so closely together that they are in contact, it seems intuitively clear that any particle, however small, while moving through the agglomerate in a straight-line motion, would rebound from one or another of the atoms before traveling very far. Even allowing for the fact that the atoms in the solid usually order into a lattice configuration, there still will be very few directions through the ordered array in which a particle could travel unimpeded on the basis of this purely classical picture. Despite this, it can be deduced from experimental measurements that under conditions of very low strain, very high purity, and quite low temperatures, the conduction electrons can travel distances involving hundreds of atoms without being scattered. Devising an acceptable explanation for such easy flow of electrons in metals thus constituted a problem that could not be resolved by means of a mechanics based on a purely classical viewpoint.

Before delving into the quantum mechanical explanation of easy electron transport in metals, let us first ask how it is known that atoms in a solid are actually in contact. Next, let us ask how it is known that electrons have such long, mean-free paths in metals for which the atoms are in contact. Classical radii for atoms may be deduced in a variety of ways. Scattering experiments initiated by Rutherford in the early 1900s provided direct evidence that an atom has a tiny, dense nucleus surrounded by a cloud of electrons extending for distances of the order of angstroms ($1\ \text{Å} = 10^{-10}$ m). The viscosity of fluids and the molecular flow of gases also provide some data on atom and molecule sizes. X-ray diffraction by crystalline solids yields lattice distances that are of the same order as the atom sizes. Compressibility data for solids lend credence to the view that forces between atoms increase as a high power of the separation distance, as might be expected from a hard-sphere picture of atoms in contact. These indications, together with a variety of other types of data, lead us to picture a solid as an array of hard-sphere atoms in contact.

The second point—namely, the existence of the long, mean-free path of electrons in metals—can be deduced from the simple picture that resistance is due to electron scattering, coupled with experimental data on the temperature-dependence of the resistivity of metals and the dependence of the resistivity on purity and crystal-preparation techniques. Perhaps the most salient point is that the resistivity decreases by many orders of magnitude when the purity of the metal is increased and the metal is grown as a strain-free single crystal.

The limiting factor on the electron mean free path in metals thus actually appears to depend upon residual imperfections, impurities, and grain boundaries in the sample instead of on atom density. There is hardly any way a classical picture can explain the fact that the ordinary atoms that make up the ordered solid themselves provide so little resistance to electron flow. The classical picture of a pointlike electron scattering in a billiard-ball manner from a hardsphere atom fails completely.

B. Quantum Mechanics Approach

Quantum mechanics permits a rationalization of the classically unexplainable observations just described. Even neglecting the ordinary Coulomb repulsion between electrons, there remains a quantum mechanical tendency for electrons to remain separated. This tendency can be treated within the framework of what is called the Pauli exclusion principle, which states that no two electrons in a system can have the same set of quantum numbers. Practically speaking, this requires higher and higher average kinetic energies for the electrons as the electron density increases. This explains why adjacent atoms resist electron-cloud overlap, even though the electron cloud otherwise would be expected to be rather soft and easily deformed under compression, and so accounts for the hard-sphere view of atoms in a crystal lattice.

The unimpeded motion of electrons moving through a lattice of such hard-sphere atoms in a solid can be understood from the wavelike properties of the electron. Even classically, it can be shown that the collective scattering of waves from a periodic array of scattering centers differs quite dramatically from the scattering of waves from a random array of scattering centers. The difference between these two situations is that a random array leads to random phases between the scattered wavefronts whereas phase coherence between the scattered wavefronts is possible if the scattering centers are located in a periodic array. (Indeed, X-ray diffraction by crystalline solids hinges on phase coherence.) In the random-array case, movement of an incident wave through the array is grossly impeded due to the partial cancellation of wavefronts having random phase with respect to one another; in the periodic array case, propagation of the wavefront becomes quite possible.

Even in the periodic case, however, there are situations in which propagation is retarded, as when a portion of the wavefront reflected from one plane of the crystalline array is superimposed upon and has a 180° phase difference with respect to another portion of the wavefront reflected from a different plane of the array. Such waves interfere destructively. Propagation, on the other hand, is enabled by a constructive interference of the scattered waves in the direction of propagation.

These facts of classical wave propagation are applicable immediately to electron propagation in solids once it is admitted that electrons have a wavelike character. Thus, it can be stated that due to the wavelike properties of electrons, the perfectly periodic array of atoms in a solid may not scatter electrons out of their straight-line path. In this sense, the periodic array may be considered to offer no resistance whatsoever to electron motion, thereby rationalizing the long, mean free paths for electrons in single, strain-free crystals of high purity held at low temperatures.

The emergent picture is that electrical resistance is not due to the scattering of electrons by the atoms of the periodic array per se but by the departures from periodicity in the crystalline array. Such departures from periodicity are provided by impurities, vacancies, strained regions, dislocations, and grain boundaries and also by thermal fluctuations of the atom array. Increased scattering at higher temperatures due to temperature-dependent thermal fluctuations in the lattice can be shown to lead to the linear temperature-dependence of the resistivity of metals. The residual resistance at extremely low temperatures is due to scattering from the impurity atoms and structural defects. A quantum mechanical approach involving the Schrödinger equation, based as it is on the wavelike behavior of particles, provides a suitable framework for rationalizing and treating these varied contributions to the electron resistivity of metals.

C. Periodic Potential for Crystalline Solids

The Schrödinger equation can be applied to describe conduction electrons in metals, each conduction electron being considered to be under the influence of a potential energy function that has the same periodicity as the lattice. An illustration of a periodic potential in one dimension is given in Fig. 9. If the lattice positions in a three-dimensional array are

$$\mathbf{R_j} = j_1\mathbf{d}_1 + j_2\mathbf{d}_2 + j_3\mathbf{d}_3, \tag{320}$$

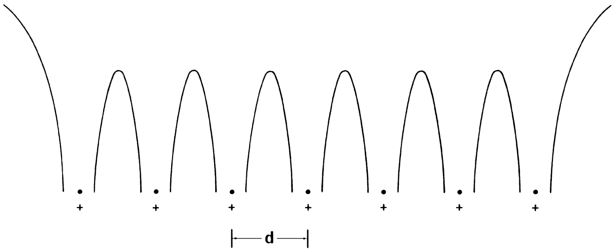


FIGURE 9 Periodic potential. [Fig. 7.8 in *Quantum Mechanics for Applied Physics and Engineering* by Albert Thomas Fromhold, Jr. (Academic Press, Inc., New York, 1981; Dover Publications, Inc., New York, 1991); reproduced with the permission of Academic Press, Dover Publications, and the author.]

where j_1, j_2, and j_3 are integers and $\mathbf{d}_1$, $\mathbf{d}_2$, and $\mathbf{d}_3$ are elemental vectors denoting the basic three units of periodicity in a three-dimensional crystalline solid, then a satisfactory potential energy $U(\mathbf{r})$, denoted by $V(\mathbf{r})$ in this special case, has the periodicity requirement

$$V(\mathbf{r}+\mathbf{R_j}) = V(\mathbf{r}). \tag{321}$$

The time-independent Schrödinger equation given in Eq. (169) then takes the form

$$-\frac{\hbar^2}{2m}\nabla^2\phi_\ell + V(\mathbf{r})\phi_\ell = \mathcal{E}_\ell\phi_\ell. \tag{322}$$

The next step is to utilize some mathematical function for $V(\mathbf{r})$. This can be assumed to be a simple form, such as a one-dimensional, periodic step function (the Krönig–Penney model), or it may be quite complex, as when one attempts to simulate mathematically the actual potential energy that would be sensed by an electron that probes different positions within each atom and between atoms in the periodic array. One very general method that lends great insight into the general problem of the motion of an electron in a periodic solid is to express the potential energy as a type of Fourier series

$$V(\mathbf{r}) = \sum_{\mathbf{n}} V_{\mathbf{n}} e^{i\mathbf{G}_\mathrm{n}\cdot\mathbf{r}} \tag{323}$$

with the amplitudes $V_{\mathbf{n}}$ for the various harmonics chosen so that the function $V(\mathbf{r})$ reproduces any periodic potential energy of interest. For a lattice in which the basic spatial periodicity vectors $\mathbf{d}_1$, $\mathbf{d}_2$, and $\mathbf{d}_3$ are orthogonal, the vectors $\mathbf{G_n}$ turn out to be especially simple in form. For the more general case of a nonorthogonal triad $\mathbf{d}_1$, $\mathbf{d}_2$, $\mathbf{d}_3$, however, it greatly facilitates the problem to define a triad $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$, called the reciprocal lattice vectors

$$\begin{aligned}\mathbf{b}_1 &= (\mathcal{V}_{\text{cell}})^{-1}\mathbf{d}_2\times\mathbf{d}_3\\ \mathbf{b}_2 &= (\mathcal{V}_{\text{cell}})^{-1}\mathbf{d}_3\times\mathbf{d}_1\\ \mathbf{b}_3 &= (\mathcal{V}_{\text{cell}})^{-1}\mathbf{d}_1\times,\mathbf{d}_2\end{aligned} \tag{324}$$

where

$$\mathcal{V}_{\text{cell}} = \mathbf{d}_1\cdot(\mathbf{d}_2\times\mathbf{d}_3) \tag{325}$$

is the volume of a unit cell in the real lattice. In terms of these vectors, the vectors $\mathbf{G_n}$ appearing in Eq. (324) are given by

$$\mathbf{G_n} = 2\pi(n_1\mathbf{b}_1 + n_2\mathbf{b}_2 + n_3\mathbf{b}_3) \tag{326}$$

where n_1, n_2, and n_3 are integers. The symbol $\mathbf{n}$ is used to represent the integer triplet n_1, n_2, n_3. The set $\mathbf{G_n}$ maps out a lattice of points in the same manner that the set $\mathbf{R_j}$ maps out a lattice, but the two lattices are generally quite different. The vectors $\mathbf{R_j}$ are said to map out a real or direct lattice, whereas the vectors $\mathbf{G_n}$ are said to map out the reciprocal lattice. The vectors $\mathbf{G_n}$ are referred to as reciprocal lattice vectors. The functions $\exp(i\mathbf{G_n}\cdot\mathbf{r})$ can be shown to have the properties required of basis functions for a Fourier series representation of arbitrary functions having the lattice periodicity.

Once the periodic potential energy is defined, then the Schrödinger equation given in Eq. (322) can be solved by various methods. One way leading to great insight into this problem is to assume a general form for the eigenfunctions ϕ_ℓ by utilizing a Fourier series description with the periodicity of the entire solid—namely,

$$\phi_{\mathbf{m}} = \sum_{\mathbf{m}} B_{\mathbf{m}} e^{i\mathbf{k}_\mathrm{m}\cdot\mathbf{r}}. \tag{327}$$

This leads to a coupled set of algebraic equations for the unknown coefficients $B_{\mathbf{m}}$ that contain the energy eigenvalues $\mathcal{E}_\ell$. The vectors $\mathbf{k_m}$ can be obtained similarly to the way the vectors $\mathbf{G_n}$ were constructed. The algebraic equations so derived constitute a homogeneous set. Self-consistency then requires the determinant of the coefficients of the unknowns $B_{\mathbf{m}}$ to be zero. This determinant, called the secular determinant, leads directly to an algebraic equation known as the secular equation for the energy eigenvalues $\mathcal{E}_\ell$. Choice of a specific eigenvalue $\mathcal{E}_\ell$ for solution of the set of algebraic equations containing the lattice potential energy coefficients $V_{\mathbf{n}}$ yields the eigenfunction ϕ_ℓ corresponding to that energy. Repeating the procedure for each energy eigenvalue, in principle, will yield the complete set of energy eigenfunctions for that periodic potential energy.

The energy eigenfunctions obtained for a periodic potential energy are known as Bloch functions, after Felix Bloch (1905–1983). Bloch functions have the general form

$$\phi_{\mathbf{m}'}(\mathbf{r}) = u_{\mathbf{m}'}(\mathbf{r})e^{i\mathbf{k}_{\mathbf{m}'}\cdot\mathbf{r}}, \tag{328}$$

where $u_{\mathbf{m}'}(\mathbf{r})$ are functions having the periodicity of the real lattice. This periodicity condition is

$$u_{\mathbf{m}'}(\mathbf{r}) = u_{\mathbf{m}'}(\mathbf{r}+\mathbf{R_j}). \tag{329}$$

The vectors $\mathbf{k}_{\mathbf{m}'}$ are propagation-type vectors for plane-wave functions having wavelengths greater than the basic unit of lattice periodicity but less than the length of the crystal in the propagation direction.

A Bloch function for an electron in a solid may be likened to an individual harmonic of sound in a musical cabinet. Bloch functions can be shown to have a number of very interesting properties, such as completeness of the set, linear independence, and orthogonality.

D. Energy Bands and Energy Gaps

The picture that thereby emerges is of groups of closely spaced, allowed discrete energies that can be populated by electrons, with the groups of allowed levels being separated by energy ranges called gaps that contain no allowed energy values for conduction electrons. Each group of closely allowed discrete energies is called an energy band. Each allowed energy value within a band is characterized by a set of quantum numbers. With the additional consideration of electron spin, these are four in number. An electron in one of the allowed levels characterized by specified values of these quantum numbers travels unscattered by the atoms of the crystal lattice, the straight-line motion being allowed due to a wavelike propagation through the spatially periodic lattice potential energy. This provides the quantum mechanical explanation of the long, mean-free path for conduction electrons in metals.

As a rule, the electrons in a solid are characterized by the vector **k**, which denotes the propagation direction. The **k**-vector is the analog of the momentum for a particle in a solid. (In free space, $\mathbf{p} = \hbar\mathbf{k}$, according to the de Broglie relation derived in Section V.B.) The energy of the electron state is denoted by $\mathscr{E}(\mathbf{k})$. The goal of solid-state band-structure calculations is to evaluate $\mathscr{E}(\mathbf{k})$ for a specified periodic potential energy. This periodic potential may be considered to be available at the outset, although the problem is best approached from the standpoint of computing the potential self-consistently with the electron states deduced in the calculation. Since $\mathscr{E} = \hbar\omega$, the function $\mathscr{E}(\mathbf{k})$ obtained by means of a band-structure calculation represents the dispersion relation for electrons in the solid. As recalled from Section IV.H, the dispersion relation provides the basis for determining the group velocity of the particle. Thus, from Eq. (107)

$$\mathbf{v}_{\text{group}} = \nabla_{\mathbf{k}}\omega(\mathbf{k}) = \frac{1}{\hbar}\nabla_{\mathbf{k}}\mathscr{E}(\mathbf{k}). \qquad (330)$$

This relation is very useful for obtaining the conduction electron velocity as a function of energy for any particular direction in the crystal. In fact, this approach must be used in lieu of the free-space relation $\mathbf{v}_{\text{particle}} = \mathbf{p}/m$ because the inertia of an electron in a crystal is governed by an effective mass m^* instead of its actual mass m. The difference in value between m^* and m is a measure of the average conduction electron interaction with the periodic potential of the lattice. Although a perfectly periodic lattice does not offer a resistance by scattering the conduction electrons, it does offer a resistance to the acceleration of the electron under an applied force (e.g., an electric field) by affecting its inertial response to the force.

In considering the population of the various energy levels, it is necessary to add in as an essential component the Pauli exclusion principle—that very soul of quantum mechanics that disallows any two electrons to occupy the same state, the state being denoted by the specification of values for the complete set of quantum numbers, including electron spin. Once this is woven into the fabric, it must be considered how the energy eigenstates are occupied by the electrons available for conduction. Under thermal equilibrium condition at temperatures near absolute zero, the lower-energy states will certainly be occupied. The lower-energy states in any of the energy bands represent states having low kinetic energies. Because the Pauli exclusion principle does not allow more than one conduction electron to crowd into any lower-energy state, higher-energy states will be populated to the degree required for all conduction electrons to be accommodated.

The requirement by quantum mechanics that all electrons be in different states has no analog in a classical mechanics description. Classically, therefore, there is no lower limit to the energy of the conduction electrons; in fact, in a classical description, the kinetic energy of the conduction electrons decreases to zero as the absolute temperature approaches zero. In a quantum mechanical description, the average kinetic energy of the conduction electrons in a metal decreases asymptotically to a still relatively high value as the absolute temperature approaches zero.

E. Metals

The average kinetic energy for the highest populated energy band can be estimated by invoking a “particle in a box” model of a metal holding its conduction electrons within the boundary walls, with no consideration given to the actual periodic potential energy of the lattice. This simple approach, known as the free-electron model, often yields surprisingly accurate quantitative values for a number of physical properties of metals associated with the conduction electrons. In such cases, a full solution of the Schrödinger equation for the actual periodic potential may not be required.

The kinetic energy of the highest filled state in a given energy band at 0 Kelvin (K) is designated the Fermi energy. A computation of how the average energy changes with increases in the thermodynamic temperature of the system yields the specific heat of the conduction electrons. The accurate predictions obtained by quantum mechanics for the specific heat of metals at low temperatures represents a remarkable success for the theory, that is to be sharply contrasted with the total failure on the part of the classical approach to provide an adequate quantitative estimate of this physical property of metals.

Quantum mechanics gives great insight into the scattering of conduction electrons by imperfections in a metal. The quantum nature of the scattering of conduction electrons places the restriction on the process that scattering can take place only to vacant quantum states of the system. This means that at 0 K, an elastic scattering event can occur only for a conduction electron having an energy equal to the Fermi energy, because that is the only energy at which both filled and empty states simultaneously exist. The situation is not quite so restrictive at higher temperatures, where there is a statistical probability that nearby states are occupied or unoccupied over a range of energy at least $k_B T$ in width in the neighborhood of the Fermi energy $\mathscr{E}_F$ [Boltzmann constant $k_B = 1.38 \times 10^{-23}$ J/K; $T =$ absolute (Kelvin) temperature]. Nevertheless, electron scattering and, hence, the electrical resistivity are still severely restricted in metals by the requirements of the Pauli exclusion principle.

F. Insulators

After accepting the preceding reasons for the success of quantum mechanics in describing the properties of the long, mean-free path in metals, one then must feel quite puzzled when confronted with the experimental fact that some crystals—even in the limit of high purity, low temperature, and perfect periodicity—do not allow the free and easy motion of electrons. These materials are called "electrical insulators."

What is it about insulators that causes them to differ so drastically from metals in the ability to conduct electrons? The answer is again quantum mechanical in origin. It is hardly more abstruse than the answer to the question of the existence of the long, mean-free path in metals. The solution of the Schrödinger equation for a periodic potential energy in the way previously outlined yields a division of the energy scale into interspersed allowable and forbidden regions of energy. Over the allowable regions (the energy bands), very closely spaced discrete energy eigenvalues are found, but within the forbidden regions (the energy gaps), there are no such energy eigenvalues. However, this property of the energy-eigenvalue spectrum characteristic of the periodic potential is, of itself, insufficient to explain the basic nature of insulators. As in the situation for electron scattering by impurities, allowance must be made for the consequences of the all-pervading Pauli exclusion principle.

A force for directed electron motion invariably leads to the prediction of a nonzero electrical current from a purely classical viewpoint; however, it does not necessarily lead to such a consequence in quantum mechanics. The reason is that acceleration of electrons by a force leads to a change in electron momentum and, generally, to an accompanying change in the electron energy. A change in electron momentum is synonymous with a change in the quantum numbers characterizing the occupied electronic state; that is, the acceleration of an electron in quantum mechanics is described by the electron vacating the state it initially occupied as it simultaneously enters a different allowed state, which, by the requisites of the Pauli exclusion principle, must necessarily be vacant before any occupation can occur. Quantum mechanically, one views the electron as being induced to enter a succession of adjacent allowed states by electric-field-induced transitions. This view can be contrasted sharply with the classical picture of a continuous acceleration of the electron through a continuous sequence of momentum vectors. For a metal having a partly filled energy band, there indeed exists the requisite sequence of nearby unoccupied states adjacent to the filled states, so that transitions can be induced by the electric force acting on the conduction electrons.

For the specific case of electrical insulators, consider the seemingly unlikely situation that there are precisely enough conduction electrons to fill every energy state up to the start of a given energy gap. Both the scattering of the conduction electrons and the electric-field excitation of electrons to nearby empty states then are impossible at 0 K and, for all practical purposes, nearly impossible even at nonzero temperatures for which $k_B T$ is far smaller than the energy gap $\mathscr{E}_{\text{gap}}$ extending to the next-higher empty energy level. Although zero scattering might seem to constitute the ideal situation leading to a low (or even zero) resistivity, there nevertheless is no existing electrical current in the presently described situation, nor can there be any induced electric current. There is no existing current because, in such a situation, all electrons occupy pairs of states representing equal magnitude but oppositely directed momentum, so that there is no net charge transport. There can be no induced current because the electric force cannot change the momentum of the electrons in any of the filled states since there are no nearby unoccupied states for the transitions. This is the required situation for an insulator.

In actuality, the seemingly unlikely situation of there being exactly enough electrons to fill a band, with none left over for the next-higher empty band, is not too unlikely. The reason for this is again a bit abstruse; in simplest terms, an energy band is found to contain one allowed state per atom in the solid for each degree of freedom of the electron spin. If there are two degrees of freedom for the electron spin (viz., two spin states, designated “up” and “down”), then the requirement for a filled energy band is simply that two conduction electrons be furnished by each atom in the crystal. Since the valence electrons become the conduction electrons in a solid, this is not an improbable condition.

G. Semiconductors

The energy-level distribution for a solid can be quite a bit more complex than just described, since there is the possibility of overlapping energy bands (i.e., energy bands unseparated by the usual energy gap). In addition, there can be energy gaps that are quite small relative to laboratory values of $k_B T$. It can be readily appreciated that overlapping bands promote metallic conduction or else can lead to what is known as a semimetal, whereas narrow energy gaps can lead to what is called a semiconductor. In semiconductors, an increase in temperature leads to more excitation of electrons from the highest-energy filled band across the gap to the adjacent empty band, thereby yielding electrons capable of conducting in what otherwise would be an empty band. Those electrons excited across the gap leave behind empty states (called electron holes) in the otherwise filled band. These empty states also can promote conduction in the following sense. Filled states nearby in momentum and energy to the newly provided empty states can undergo transitions to the empty states by means of electric-field excitation, all such transitions being impossible in the 0 K equilibrium situation where all states in the band are filled. In this way, two distinct carrier types are simultaneously provided—namely, electrons in an almost empty band and electron holes in an almost filled band.

The number of electrons excited across the energy gap increases nearly exponentially with increasing temperature. To the extent that electron transport increases with the number of carriers, the conductivity of the semiconductor increases almost exponentially with the temperature. A material having the properties just described is called an intrinsic semiconductor.

The prediction of the experimentally observed exponential increase of conductivity with increasing temperature in intrinsic semiconductors represents another triumph of quantum mechanics. In a classical description, there are no energy gaps and, hence, no parallel to predictions of an exponentially increasing conductivity due to excitation across an energy gap.

The exponential increase of conductivity with temperature in semiconductors also contrasts markedly with the temperature-dependence of the conductivity of metals. In that case, the conductivity decreases (instead of increasing) with increasing temperature. This decrease of conductivity in metals, which is more or less linear with increasing temperature, is due to the increase in the thermal vibrations of the atoms of the lattice at higher temperatures. Thermal vibrations yield greater departures of the atom array from perfact periodicity, thereby leading to more random scattering of the electrons and a consequent decrease in the electron current.

The exponential increase in conductivity with temperature described for excitation of electrons across an energy gap in an intrinsic semiconductor is paralleled in another type of solid, called an extrinsic semiconductor. In extrinsic semiconductors, however, there is no band-to-band excitation. Instead, the source of electrons for the empty band is a doping concentration of impurities that have outer electrons at energies just below the empty band (so-called donor impurities, or simply donors) or, alternately, empty levels at energies just above the filled band (so-called acceptor impurities, or simply acceptors). Semiconduction then takes place by means of electrons donated to the empty band by donor impurities or, alternately, by electron holes created in the filled band as a result of electrons accepted from that band by the acceptor impurities.

H. Superconductors

It is another well-known experimental fact that a number of metals go into a state of zero resistance at very low temperatures, below the so-called transition temperature characteristic of the material in question. The concept of energy gaps likewise turns out to play an important role in understanding this incredible phenomenon, although in a different way from that just described for semiconductors. The energy gap in the case of superconductors is attributed to a condensation of the electrons carrying the charge into so-called Cooper pairs, the binding energy of the pairs being attributed to indirect Coulomb-force-induced interaction between electrons as mediated by the intervening ion cores on the lattice sites in the metal.

Let us consider, for example, pairs of electrons passing one another while traveling in opposite directions through the lattice of ion cores surrounded by the attendant electron clouds. One electron exerts a force on the nearby ionic lattice, which responds to that force. The resulting disturbance in the periodic potential sensed by the second electron of the pair can lead to an effective lowering in the total energy relative to the situation of a rigid, nonresponding lattice. The energy lowering is the greatest for pairs of electrons having equal magnitudes but oppositely directed momentum values, so Cooper pairs are characterized by two electrons having this property. The lowering of energy leads to a superconducting energy gap. The energy gap so produced is quite small, so that very low temperatures are usually required for the pairs to remain unbroken by thermal fluctuations. The temperature range of superconductivity has been greatly extended, however, by the discovery of the so-called “high temperature superconductors” initiated with a series of compounds based on a copper oxide matrix.

As the temperature is reduced through the superconducting transition, one speaks of the condensation of the electron system into the paired state. It is evident that electrons with oppositely directed momentum values will become separated spatially in a short time, so the pairing process must be statistical in nature. Pairs must continually exchange partners (as required, e.g., in some forms of folk dancing).

The dance of the conduction electrons, while maintaining this property of electron pairing, is a many-body problem of some complexity. The wavefunction for the entire system of electrons as a unit must be considered, not merely the single-particle wavefunctions individually. The establishment of a net electric current requires a suitable modification in the zero-current wavefunction for the system.

The importance of electron pairing for electrical resistance is that the scattering of conduction electrons in the paired state will be ineffective in randomizing the net electron momentum. Thus, there can be a zero resistivity state as long as Cooper pairs exist in the system.

Thermal fluctuations can break Cooper pairs to yield electrons in the normal state. The breaking of electron pairs by this means is tantamount to an excitation across the energy gap. In contrast to the case of semiconductors, the excitation across the energy gap in the present instance leads to an increase in the resistivity. Raising the temperature of a superconductor through the superconducting transition temperature means that the thermal fluctuations become so large that essentially no Cooper pairs remain in the metal to provide superconductivity.

I. Success of Quantum Mechanics for Electron Transport in Solids

Thus, quantum mechanics provides a framework for understanding the widely different electrical-conduction properties of superconductors, normal metals, semiconductors, and insulators. This remarkable success, coupled with the parallel failures of classical physics to lend understanding to these areas, has led to the nearly universal acceptance of quantum mechanics for most calculations in solid-state physics.

X. SUMMARY

The theory of quantum mechanics evolved in the 1920s to correlate and predict the behavior of atomic and subatomic systems. Heisenberg and Schrödinger played prominent roles in the development of this theory, which proves to be the only formulation adequate for the microscopic domain of nature. Heisenberg stressed the importance of including physical observables and experimental observations of optical spectral lines in his formulation. Schrödinger based his work on a differential equation for the wavelike behavior of small mass particles that includes the possibility of constructive and destructive interference of waves presumably associated with the presence of a particle. Inherent in quantum mechanics is the germ concept that accurate predictions of future trajectories of particles and the time evolution of a system involving one or more particles are at best statistical, involving a range of possibilities specified exactly only in terms of precise values for the relative probabilities. This precludes the deterministic prediction of an exactly specified future path, regardless of how accurately the initial conditions of the system are specified. Also inherent in the theory is the impossibility of exactly measuring even the initial conditions, such as initial position and initial linear momentum, due to some inherent uncertainty in the value of one of these variables following determinations of the values of the other variables to some specified degree of precision. The philosophical implications of an inherent uncertainty in quantum mechanical predictions, contrasted with the absolute determinism inherent in classical mechanical predictions, initially led many (including Einstein) to doubt whether the discipline had any fundamental merit beyond that of being an elegant and elaborate computation tool for obtaining predictions of a statistical nature. To date, no better theory for the microscopic world has been developed.

SEE ALSO THE FOLLOWING ARTICLES

Bonding and Structure in Solids • Celestial Mechanics • Electrodynamics, Quantum • Electromagnetics • Lasers • Mechanics, Classical • Particle Physics, Elementary • Quantum Chemistry • Quantum Optics • Relativity, General • Relativity, Special • Statistical Mechanics

BIBLIOGRAPHY

Bohm, D. (1951). “Quantum Theory,” Prentice-Hall, Englewood Cliffs, NJ.

Born, M. (1969), “Physics in My Generation,” Springer-Verlag, New York.

de Broglie, L. (1939). “Matter and Light,” Dover Publications, New York.

Fromhold, A. T., Jr. (1981, 1991). “Quantum Mechanics for Applied

Physics and Engineering," Academic Press, New York; Dover Publications, New York.

Heisenberg, W. (1930). "Physical Principles of Quantum Theory," Dover Publications, New York.

Ikenberry, E. (1962). "Quantum Mechanics for Mathematicians and Physicists," Oxford University Press, New York.

Mandl, F. (1957). "Quantum Mechanics," Butterworths Scientific Publications, Stoneham, MA.

Pauli, W. (1973). "Wave Mechanics," Pauli Lectures on Physics, Vol. 5. MIT Press, Cambridge, MA.

Pauling, L., and Wilson, E. B., Jr. (1935). "Introduction to Quantum Mechanics," McGraw-Hill, New York.

Quantum Optics

J. H. Eberly
University of Rochester

P. W. Milonni
Los Alamos National Laboratory

GLOSSARY

AC Stark effect Effective increase of the Bohr transition frequency of a two-level atom which is being excited by a strong laser beam, the amount of increase being the Rabi frequency.

Bloch vector Fictitious vector whose rotations are equivalent to the time dependence of the wave function or quantum mechanical density matrix associated with a two-level atom.

Coherence time Limiting time interval between two segments of a light beam beyond which the superposition of the segments will no longer lead to interference fringes.

Coherent state Quantized state of a light field whose fluctuation properties are Poissonian; it is considered the most classical quantized field state.

Degree of coherence Normalized measure of the ability of a light beam to form interference fringes.

Optical bistability Existence of two stable output intensities for a given input intensity of a steady light beam transmitted through a nonlinear optical material.

Optical Bloch equations Dynamical equations that determine the motion of the Bloch vector; they are a special type of quantum Liouville equation.

Photon echo Burst of light emitted by a collection of two-level atoms signaling the realignment of their Bloch vectors after initial dephasing; similar to the spin echo of nuclear magnetic resonance.

Rabi frequency Steady frequency of rotation of the Bloch vector of an atom exposed to a constant laser beam, proportional to the atom's transition dipole moment and the laser's electric field strength.

Superradiance Spontaneous emission from many atoms exhibiting collective phase-coherence properties, such as radiation intensity proportional to the square of the number of participating atoms.

Two-level atom Fictitious atom having only two energy levels which is used as a model in theoretical studies of near-resonance interactions of atoms and light, particularly laser light.

QUANTUM OPTICS is the study of the statistical and dynamical aspects of the interaction of matter and light. It is concerned with phenomena ranging from spontaneous emission and single photon absorption to the highly nonlinear processes induced by laser fields and has connections with laser physics, nonlinear optics, quantum electronics, quantum statistics, and quantum electrodynamics.

I. INTRODUCTION

A. Central Issues of Quantum Optics

Planck's quantum, announced to the Prussian Academy on October 19, 1900, as a solution to the blackbody puzzle reopened the wave–particle question in optics, a question that Fresnel and Young had settled in favor of waves almost two centuries earlier. Planck's quantum could not be confined to light fields. Within three decades, all of particle mechanics had been quantized and rewritten in wave mechanical form, and wave–particle duality was understood to be both universal and probabilistic.

Quantum optics is fundamentally concerned with coherence and interference of both photons and atomic probability amplitudes. For example, it provides one of the main avenues at the present time for detailed study of wave–particle duality. The central issues of quantum optics deal with light itself, with quantum mechanical states of matter excited by light, and with the process of interaction of light and matter.

Questions arising in the description of a single atom and its associated radiation field, as the atom makes a transition between two energy states and either emits or absorbs a photon, are among the most central questions in quantum optics. Observations of individual optical emission and absorption events are possible, and the interpretation of such observations is at the heart of quantum theory.

Various elements of these central considerations are to a degree independent of each other and are understood separately. Among these are (1) the probability that an atomic electron occupies one or another state and the rate at which these occupation probabilities change, (2) the statistical nature of the photons emitted during transitions, (3) correlations between atom and photon states, (4) the characteristic parameters that control the light–matter interaction, and (5) the intrinsically quantum mechanical features of the atom's response to the radiation.

Quantum optics also concerns itself with problems that grow out of these central considerations and whose answers can be expressed within the conceptual framework established by the central problem. Areas related in this way to the core of quantum optics deal, for example, with correlated many-atom light–matter interactions; near-resonant transitions among three and more states of an atom or molecule or solid; optical tests of quantum electrodynamics and measurement theory; multiphoton processes; quantum limits to noise and linewidth; quantum theory of light amplification and laser action; and manifestations of nonlinearity, bistability, and chaos in optical contexts. A wide variety of quantum optical phenomena that bear on one or another of these issues are now known and widely studied.

B. The *A* and *B* Coefficients of Einstein

The second half of the twentieth century saw remarkable advances in our understanding of light, of its generation, propagation, and detection. The laser is one manifestation of these advances. Lasers generally depend on the quantum mechanical properties of atoms, molecules, and solids because quantum properties determine the ways that matter absorbs light and emits light. Conversely, the properties of laser beams have made optical studies of quantum mechanics possible in a variety of new ways. It is this interplay that has created the field of quantum optics since about 1960.

From a different historical perspective, however, quantum optics is much older than the laser and even older than quantum mechanics. The quantum concept first entered physics in 1900 when Planck invented the light quantum to help understand black body radiation. The understanding of other quantum optical phenomena, such as the photoelectric effect, first explained by Einstein in 1905, was well underway almost two decades before a quantum theory of mechanics was properly formulated in 1925 and 1926 by Heisenberg and Schrödinger. Indeed, these early developments in quantum optics played an essential role in the first quantum pictures of atomic matter given by Bohr and others in the period from 1913 to 1923.

Only two parameters are needed to understand the interaction of light with atomic (and molecular) matter, according to Einstein. These two parameters are called A and B coefficients. These coefficients are important because they control the rates of photon emission and absorption processes in atoms, as follows. Let the probability that a given atom is in its nth energy level be written P_n. Suppose there are photons present in the form of radiation with spectral energy density (J/m^3 Hz) denoted by $u(\omega)$. Then the rate at which the probability P_n changes is due to three fundamental processes:

$$(dP_n/dt)_{\text{absorption of light}} = +Bu(\omega)P_m \tag{1a}$$

$$(dP_n/dt)_{\text{spontaneous emission of light}} = -AP_n \tag{1b}$$

$$(dP_n/dt)_{\text{stimulated emission of light}} = -Bu(\omega)P_n. \tag{1c}$$

Here P_m is the probability that the atom is in a lower level, the mth, which is related to the nth through the energy relation $E_n - E_m = \hbar\omega$, where $\hbar = h/2\pi$ and h is Planck's famous quantum constant. Einstein's great insight was to include stimulated emission [Eq. (1c)] among the three elementary processes, in effect, to recognize that an atom in an upper energy state could be encouraged by the presence of photons [the existence of $u(\omega)$] to hasten the rate at which it would drop down to a lower state.

The three contributions to the rate of change of P_n shown in Eqs. (1a–c) can be added to make an overall single equation for the total rate of change of P_n:

$$dP_n/dt = +Bu(\omega)P_m - AP_n - Bu(\omega)P_n. \tag{2a}$$

Einstein applied this equation to an examination of blackbody light. He showed that the steady-state solution

$$P_m/P_n = 1 + A/Bu(\omega) \tag{2b}$$

implies the validity of Planck's formula for $u(\omega)$:

$$u(\omega, T) = \frac{\hbar\omega^3}{\pi^2 c^3}\frac{1}{\exp\{\hbar\omega/kT\} - 1}, \tag{3}$$

and the value of the prefactor is just the ratio A/B:

$$\frac{A}{B} = \frac{\hbar\omega^3}{\pi^2 c^3}, \tag{4}$$

where k is Boltzmann's constant and T the temperature in degrees Kelvin. Table I contains the values of physical constants used in evaluating various radiation formulas. For typical optical radiation, the value of the fundamental ratio A/B is approximately 10^{-14} J/m^3 Hz. The corresponding intensity, namely cA/B, is approximately 3×10^{-6} J/m^2, or $6\pi \times 10^{-6}$ W/m^2 per Hz of bandwidth. The value of the spectral intensity of thermal radiation is usually many orders of magnitude lower than this because of the second factor in Eq. (3). At optical wavelengths, the second factor is much smaller than one for all temperatures less than about 5000 K.

After the development of a fully quantum mechanical theory of light by Dirac in 1927, it was possible to give expressions for A and B separately:

$$A = \frac{1}{4\pi\varepsilon_0}\frac{4D^2\omega^3}{3\hbar c^3} \tag{5}$$

$$B = \frac{1}{4\pi\varepsilon_0}\frac{4\pi^2 D^2}{3\hbar^2}. \tag{6}$$

TABLE I Physical Constants Used in Evaluating Radiation Formulas

Constant	Value
h (Planck's constant)	6.6×10^{-34} J/s
k (Boltzmann's constant)	1.38×10^{-23} J/K
c (Speed of light)	3×10^{8} m/s
e (Electric charge)	1.6×10^{-19} C
λ (Typical optical wavelength, yellow)	600×10^{-9} m
ν (Typical optical frequency)	5×10^{14} Hz

In these formulas we have separated the factor $1/4\pi\varepsilon_0 = 8.9874 \times 10^9$ N m^2/C^2 to display A and B in atomic units as well as SI units, and D denotes the quantum mechanical "dipole matrix element" associated with the $m \to n$ transition under consideration.

The values of these important coefficients can be obtained for transitions of interest in quantum optics by assuming that the dipole matrix element is approximately equal to the product of the electron's charge and a "typical" electron displacement from the nucleus. Thus, we take $\omega = 2\pi\nu$ and e and h from Table I and $D = er$, with r equal to about 1 to 3 Å (1–3×10^{-10} m). In this case the values are $A \approx 10^8$ s^{-1} and $B \approx 10^{22}$ m^2/J s^2, respectively.

The advantage of Einstein's approach, and the reason it still provides one basis for understanding light–matter interactions, is that it breaks the interaction process into its separate elements, as identified above in Eqs. (1a–c). To repeat this important identification, these processes are (a) absorption, (b) spontaneous emission, and (c) stimulated emission.

However, it must be pointed out that Einstein's formulas are not universally valid, and Eqs. (1) and (2) can be seriously misleading in some cases, particularly for laser light. Laser light typically has a very high spectral energy density $u(\omega)$. In this case different formulas and equations, and even entirely different concepts with their origins in wave mechanics, may be required.

A large body of experimental evidence has accumulated since 1960 showing that many aspects of the interaction between light and matter depend on electric radiation field strength E directly, not only on energy density $u \approx E^2$. Just those aspects of the light–matter interaction that depend directly on E also depend directly on quantum mechanical state amplitudes ψ, not only on their associated probabilities $|\psi|^2$. Issues of coherence and interference of both radiation fields and probability amplitudes are fundamental to these studies. It is principally the experiments and theories that deal with light and matter in this domain that make up the field of quantum optics.

C. Two-State Atom and Maxwell–Bloch Equations

As Einstein's arguments suggest, in quantum optics it is often sufficient to focus attention on just two energy levels of an atom—the two levels that are closest to resonance with the radiation, satisfying the energy condition

$$E_2 - E_1 \approx \hbar\omega,$$

where $\omega = 2\pi\nu$ is the angular frequency of the radiation field. This is shown schematically in Fig. 1.

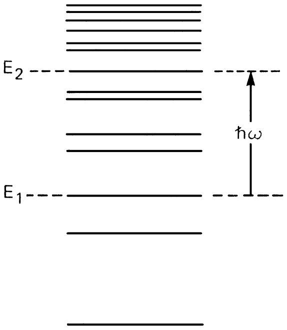


FIGURE 1 Schematic energy level diagram showing a two-level subsystem.

Under these circumstances the wave function of the atom is a sum of the wave functions for the two states

$$\psi(\mathbf{r}, t) = C_1\phi_1(\mathbf{r}) + C_2\phi_2(\mathbf{r}). \tag{7a}$$

For simplicity of description we will assume each level corresponds to a single quantum state and will usually use "level" and "state" synonymously.

The assumption that the electron is certainly in one or the other or a combination of these two levels is expressed mathematically by the equality

$$|C_1|^2 + |C_2|^2 = 1. \tag{7b}$$

Each term in this equation is called a level probability, and all probabilities must add to 1, of course. These probabilities are the quantities labeled by the letter P in the Einstein equations. The C's themselves are not probabilities, and they have no counterpart in classical probability theory. They are called probability amplitudes. It is the remarkable nature of quantum mechanics that the fundamental equation (the Schrödinger equation) governs these amplitudes C, not the probabilities $|C|^2$.

The Schrödinger equation for either one of the amplitudes is

$$i\hbar\, dC_m/dt = E_m C_m + V_{mn} C_n, \tag{8}$$

where m and n take either the value 1 or 2, but $m \neq n$. Here $\hbar$ is the usual abbreviation for $h/2\pi$, and V_{mn} is called the interaction matrix element between the atom and the radiation field. In almost all cases of interest in quantum optics, this interaction comes from the potential energy $-\mathbf{d} \cdot \mathbf{E}(\mathbf{r}, t)$ of the atomic dipole in the electric radiation field (a dipole $\mathbf{d} = e\mathbf{r}$ exists because of the separation of the negative electronic charge in its planetary orbit from the position of the positively charged nucleus.

In principle $\mathbf{E}$ is a quantum operator field (see Section III.C). The so-called semiclassical theory of radiation uses its average (or "expectation") value instead. This approximation is usually justified when the field is intense, because quantum fluctuations, which almost always occur at the single-photon level, are then negligible. In this section we describe the semiclassical theory and ignore quantum field effects entirely.

The dipole interaction is distributed over the atomic orbitals involved, with the result that

$$\begin{aligned} V_{mn} &= \langle\phi_m| - e\mathbf{r} \cdot \mathbf{E}(\mathbf{r}, t)|\phi_n\rangle \\ &\approx \int d^3r\,\phi_m^*(\mathbf{r})[-e\mathbf{r} \cdot \hat{\mathbf{e}}E(\mathbf{r}_N, t)]\phi_n(\mathbf{r}) \qquad (9a) \\ &= -\mathbf{d}_{\mathrm{mn}} \cdot \hat{\mathbf{e}}E(\mathbf{r}_N, t). \end{aligned}$$

In Eq. (9a) we have written

$$\mathbf{E}(\mathbf{r}, t) = \hat{\mathbf{e}}E(\mathbf{r}, t) \approx \hat{\mathbf{e}}E(\mathbf{r}_N, t),$$

where $\hat{\mathbf{e}}$ is the polarization and $E(\mathbf{r}_N, t)$ is the amplitude of the electric field at the position of the atom (i.e., of the nucleus) $\mathbf{r}_N$ instead of at the position of the electron. This approximation is usually well justified because the electron's orbit, and thus the range of the integral, extends mainly over a region much smaller than an optical wavelength. Thus, over the whole range of the integral $E(\mathbf{r}, t) \approx E(\mathbf{r}_N, t)$ and the only $\mathbf{r}$ dependence comes from the dipole moment $e\mathbf{r}$ itself. This is called the dipole approximation. The integral in Eq. (9a) is called the dipole matrix element, d,

$$d = \hat{\mathbf{e}} \cdot d_{mn} = \hat{\mathbf{e}} \cdot \int d^3r\,\phi_m^*(\mathbf{r})e\mathbf{r}\phi_n(\mathbf{r}). \tag{9b}$$

If only one atom is under consideration it is common to put it at the origin of coordinates and write $E(0, t)$ or simply $E(t)$. If several or many atoms are under consideration this is generally not possible [see, e.g., Eq. (20)]. Equation (8) can be written in a simpler form by anticipating an interaction with a quasi-monochromatic radiation field

$$E(t) = \mathscr{E}_0(t)e^{-i\omega t} + \text{c.c.} \tag{10}$$

that is nearly resonant with the atom, where c.c. means complex conjugate. This means that the angular frequency $\omega = 2\pi\nu$ of the radiation field is approximately equal to the angular transition frequency of the atom: $\omega_{21} = (E_2 - E_1)/\hbar$. In this case there is a strong synchronous response of the atom to the radiation, and the equations simplify if one removes the synchronously "driven" component of the atomic response by defining new variables a_n:

$$a_1 = C_1 \tag{11a}$$

$$a_2 = C_2 e^{i\omega t}. \tag{11b}$$

Note that $|a_1|^2 + |a_2|^2 = |C_1|^2 + |C_2|^2 = 1$; thus, the a's are also probability amplitudes.

If $E(t)$ is sufficiently monochromatic (the field amplitude, i.e., $\mathscr{E}_0$ is practically constant in time; which means $|d\mathscr{E}_0/dt| \ll \omega|\mathscr{E}_0|$), then a rotating wave approximation

(RWA) is valid if the anticipated atom field resonance is sufficiently sharp, that is, if $|\omega_{21} - \omega| \ll \omega$. The most important consequence of the RWA is that factors such as $1 + e^{\pm 2i\omega t}$, which appear in the exact equations for the new amplitudes a_1 and a_2, may be replaced by 1 to an excellent approximation. The result is that the frequencies ω_{21} and ω individually play no further role in the Schrödinger equation, and Eq. (8) takes the extremely compact RWA form:

$$i\, da_1/dt = -\tfrac{1}{2}\chi a_2 \tag{12a}$$

$$i\, da_2/dt = \Delta a_2 - \tfrac{1}{2}\chi a_1, \tag{12b}$$

where

$$\Delta = \omega_{21} - \omega \tag{12c}$$

is called the detuning of the atomic transition frequency from the radiation frequency, and

$$\chi = 2d\mathscr{E}_0/\hbar \tag{12d}$$

is called the Rabi frequency of the interaction. For simplicity, the Rabi frequency χ will be assumed here to be a real number.

For a strictly monochromatic field (time-independent $\mathscr{E}_0$) the solution to Eqs. (12a and b) is easily found in terms of $\Omega = \sqrt{\chi^2 + \Delta^2}$, where Ω is called the generalized or detuning-dependent Rabi frequency. In the most important single case the atom is in the lower state at the time the interaction begins, which means that $a_1(0) = 1$ and $a_2(0) = 0$. In this case the solution is

$$a_1(t) = [\cos \Omega t/2 + (i\Delta/\Omega)\sin \Omega t/2]e^{-i\Delta t/2} \tag{13a}$$

$$a_2(t) = i[(\chi/\Omega)\sin \Omega t/2]e^{-i\Delta t/2} \tag{13b}$$

and the corresponding probabilities are

$$P_1 = \cos^2(\Omega t/2) + (\Delta/\Omega)^2 \sin^2(\Omega t/2) \tag{14a}$$

$$P_2 = (\chi/\Omega)^2 \sin^2(\Omega t/2). \tag{14b}$$

These solutions describe continuing oscillation of two-level probability between levels 1 and 2. They have no steady state. Figure 2 shows graphs of $P_2(t)$. One already sees, therefore, that the quantum amplitude equations [Eqs. (12a–d)] make strikingly different predictions from the two-level equations of Einstein [recall the steady-state solution Eq. (2b)].

Within the RWA, the dynamics remain unitary or probability conserving: $P_1 + P_2 = 1$ for all values of t. As a result, the compact version [Eq. (12)] of Schrödinger's equation remains valid even for very strong interactions between the atom and the radiation. This is important when considering the effects on atoms of very intense laser fields. The limits of validity of Eq. (12) are determined to a large extent by a generalized statement of the limits of the RWA:

$$\omega_{21} \text{ and } \omega \gg \sqrt{\Delta^2 + \chi^2}. \tag{15}$$

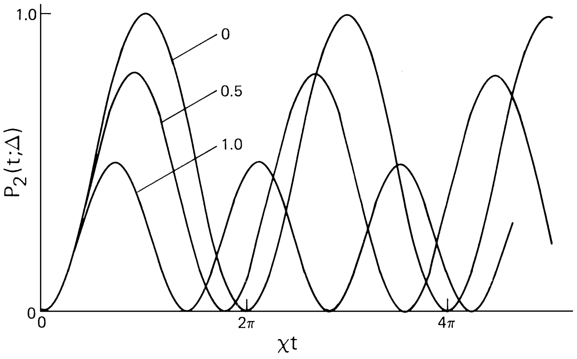


FIGURE 2 Rabi oscillations for different values of Δ/x .

There are two other real quantities associated with the amplitudes a_1 and a_2 of the radiation–atom interaction. They belong to the atomic dipole's expectation value $\langle \mathbf{d} \rangle$ which, according to Eqs. (7a), (9b), and (11), is

$$\langle \psi(t)|e\mathbf{r}|\psi(t)\rangle = a_1(t)a_2^*(t)e^{i\omega t}\mathbf{d}_{21} + \text{c.c.} \tag{16}$$

The new quantities are designated by u and v:

$$u = a_1^* a_2 + a_1 a_2^* \tag{17a}$$

$$v = i\left(a_1^* a_2 - a_1 a_2^*\right). \tag{17b}$$

Along with u and v, a third variable, the atomic inversion $w = P_2 - P_1$ plays an important role. The solutions for u, v, and w corresponding to Eqs. (13) and (14) above are

$$u = (\Delta\chi/\Omega^2)(1 - \cos \Omega t) \tag{18a}$$

$$v = (\chi/\Omega)\sin \Omega t \tag{18b}$$

$$w = -(1/\Omega^2)(\Delta^2 + \chi^2 \cos \Omega t), \tag{18c}$$

which share the oscillatory properties of the probabilities. They also obey the important conservation law:

$$u^2 + v^2 + w^2 = 1, \tag{18d}$$

which is the same as $|a_1|^2 + |a_2|^2 = 1$.

For many purposes in quantum optics the semiclassical dipole variables u and v, and the atomic inversion w, are the primary atomic variables. They obey equations which are equivalent to Eq. (8) and take the place of Schrödinger's equation for the level's probability amplitudes a_1 and a_2 (or C_1 and C_2):

$$du/dt = -\Delta v \tag{19a}$$

$$dv/dt = \Delta u + \chi w \tag{19b}$$

$$dw/dt = -\chi v. \tag{19c}$$

All of these considerations assume that the field is monochromatic or nearly so.

Although the field $\mathbf{E}(\mathbf{r}, t)$ is not considered an operator in the semiclassical formulation, it can still have a dynamical character, which means it is not prescribed in advance, but obeys its own equation of motion. It is naturally taken to obey the Maxwell wave equation, with the usual source term $\mu_0 d^2/dt^2 \mathbf{P}(z, t)$. In the semiclassical theory $\mathbf{P}(z, t) = N\langle \mathbf{d} \rangle$, where N is the density of two-level atoms in the source volume and $\langle \mathbf{d} \rangle$ the average dipole moment of a single atom, already calculated in Eq. (16).

As Eq. (16) indicates, $\langle \mathbf{d} \rangle$ is quasi-monochromatic, essentially because a two-level atom is characterized by a single transition frequency. Therefore the $\mathbf{E}$ that it generates may also be regarded as monochromatic or nearly so, as Eq. (10) assumed. This internal consistency is an important consideration in the semiclassical theory. In many cases it is also suitable to regard the field as having a definite direction of propagation, say z, and to neglect its dependence on x and y (plane wave approximation):

$$\mathbf{E}(z, t) = \hat{\boldsymbol{e}}\mathscr{E}(z, t)e^{-i(\omega t - kz)} + \text{c.c.}, \qquad (20)$$

where $k = \omega/c$ and the amplitude or envelope function $\mathscr{E}(z, t)$ is a complex generalization of $\mathscr{E}_0(t)$ in Eq. (10). It obeys a "reduced" wave equation in variables z and t:

$$[\partial/\partial z + \partial/\partial ct]\mathscr{E}(z, t) = i(\pi N d\omega/4\pi\varepsilon_0 c)[u - iv]. \qquad (21)$$

The reduced wave equation Eq. (21) and Eqs. (19) for u, v, and w are called the semiclassical coupled *Maxwell–Bloch equations*.

Inspection of the semiclassical Eqs. (19) and (21) reveals their major flaw. One easily sees that the semiclassical approach to radiation theory does not include the process of spontaneous emission. A completely excited atom will not emit a photon in this theory. That is, if the atom is in its excited state, then $a_2 = 1$ and $a_1 = 0$, so $w = 1$ and $u = v = 0$. Thus, according to Eq. (21) no field can be generated. By the same token, if $\mathscr{E} = 0$ then $\chi = 0$ and $dw/dt = 0$, and no evolution toward the ground state can occur. The flaw in the semiclassical Maxwell–Bloch approach arises from the assumption that all of the dynamics can be reduced to a consideration of average values, that is, averages of dipole moment and inversion and field strength. In reality, fluctuations about average values are an important ingredient of quantum theory and essential for spontaneous emission.

Spontaneous emission does not play a dominant role in many quantum optical processes, particularly those involving strong radiation fields. The semiclassical Maxwell–Bloch equations give an entirely satisfactory explanation of these effects, and the next sections describe some of them.

In nature there are no actual two-level atoms, of course. However, the selection rules for allowed optical dipole transitions are sufficiently restrictive, and optical resonances can be sufficiently sharp that very good approximations to two-level atoms can be found in nature. A good example is found in a pair of levels in atomic sodium, and they have been used in quantum optical experiments.

In sodium the nucleus has spin $I = 3/2$, and the lowest electronic energy level is $3S_{1/2}$ with hyperfine splitting ($F = 1$ and $F = 2$). The first excited levels $3P_{1/2}$ and $3P_{3/2}$ are responsible for the well-known strong sodium D lines of Fraunhofer in the yellow region of the optical spectrum at wavelengths 589.0 nm and 589.6 nm. The "two-level" transition is between the $m_F = +2$ magnetic sublevel of $F = 2$ of the ground state and the $m_F = +3$ magnetic sublevel of $F = 3$ in the $3P_{3/2}$ excited state. Circularly polarized dye laser light can be tuned within the 15-MHz natural linewidth of the upper level, and the $\Delta m = +1$ selection rule for circular polarization prevents excitation of the other magnetic sublevels.

Spontaneous decay from the upper state, which obeys no resonance condition, is also restricted in this example. The final state of spontaneous decay could, in principle, have $m_F = 4, 3, 2$. However, in sodium there are no $m_F = 4$ or 3 states below the $3P_{3/2}$ state, and only one $m_F = 2$ state, namely the one from which the excitation process began. Thus, the state of the sodium atom is very effectively constrained to this two-state subset out of the infinitely many quantum states of the atom.

II. INDUCED ATOMIC COHERENCE EFFECTS

The ability of an atomic system to have a coherent dipole moment during an extended interaction with a radiation field is a necessary condition for a wide variety of effects associated with quantum optical resonance. The dipole moment should be coherent in the sense that it retains a stable phase relationship with the radiation field. Long-term phase memory may be difficult to achieve, for example, because spontaneous emission and collisions destroy phase memory at a rate that is typically in the range $10^8\ \text{s}^{-1}$ or much greater. Coherence is also lost if the radiation bandwidth is too broad. The importance of dipole coherence effects shows that light–matter interactions do depend most fundamentally on the dipole moment and electric field strength, not on radiation intensity and the B coefficient.

A. p Pulses and Pulse Area

The solution for the level probabilities given in Eq. (14) is the single most important example in quantum optics of the coherent response of an atom to a monochromatic radiation field. "Coherence" in this context has several connected meanings, all associated with the well-phased steady oscillation of the probabilities considered as a function of time. This time dependence was shown in Fig. 2 for several values of the parameters Δ and χ. The significance of the Rabi frequency χ is clear—it is the frequency at which the inversion oscillates when the atom and the radiation are at exact resonance, when $\Delta = 0$, since

$$w(t) = -\cos \chi t. \tag{22}$$

Fruitful connections to the spin vector formalism of magnetic resonance physics are obtained by regarding the triplet $[u, v, w]$ as a vector $\mathbf{S}$. Equations (19) for u, v, and w can then be written in compact vector form:

$$d\mathbf{S}/dt = \mathbf{Q} \times \mathbf{S}, \tag{23}$$

where $\mathbf{Q}$ is the vector of length Ω with components $[-\chi, 0, \Delta]$. The vector $\mathbf{Q}$ can be called the torque vector for $\mathbf{S}$, which is variously called the pseudo-spin vector, the atomic coherence vector, and the optical Bloch vector.

This vector formulation in Eq. (23) of Eqs. (19) shows that the evolution of the two-level atom in the presence of radiation is simply a rotation in a three-dimensional space. The space is only mathematical in quantum optics because the components of the optical Bloch vector are not the components of a single real physical vector, whereas in magnetic resonance they are the components of a real magnetic moment. In both cases the nature of the torque equation leads to a useful conservation law: $d/dt(\mathbf{S} \cdot \mathbf{S}) = 0$. That is, the length of the Bloch vector is constant. In the u, v, w notation this means $u^2 + v^2 + w^2 = 1$, and it implies that the vector $[u, v, w]$ traces out a path on a unit sphere as the two-level atom changes its quantum state.

Further consideration of the on-resonant atoms ($\Delta \to 0$ and $\Omega \to \chi$) gives information about the interaction of atoms with (nonmonochromatic) pulsed fields. According to Eq. (19a) $u(t)$ can be neglected if $\Delta = 0$ and a new form of solutions to Eqs. (19b and c) follows immediately:

$$v(t) = -\sin \phi(t) \tag{24a}$$

$$w(t) = -\cos \phi(t), \tag{24b}$$

where $\phi(t)$ is called the "area" of the electric field because it is related to the time integral of the electric field envelope:

$$\phi(t) = \int_0^t \chi(t')\, dt' = (d/\hbar) \int_0^t \mathscr{E}_0(t')\, dt'. \tag{25}$$

In the monochromatic limit when $\mathscr{E}_0 = \text{constant}$, then $\phi(t) \to \chi t$.

Recall that the rotating wave approximation (RWA) will not permit $\mathscr{E}_0$ to vary too rapidly. If τ_p is the pulse length, then $\mathscr{E}_0$ is not too rapidly varying if τ_p is long enough, namely if $1/\tau_\mathrm{p} \ll \omega$. Pulse lengths in the range 1 ns $\geq \tau_\mathrm{p}$ $\geq$10 fs are of interest and are compatible with the RWA [1 fs (femtosecond) $= 10^{-15}$ s].

The significance of $\phi(t)$ is evident in Eq. (24). The term ϕ is just the angle of rotation of the on-resonance Bloch vector $\mathbf{S} = [u, v, w]$ during the passage of the light pulse. A light pulse with $\phi = \pi$ is called a π *pulse*, that is, a pulse that rotates the initial vector $[0, 0, -1]$, which points down, through 180° to the final vector $[0, 0, +1]$, which points up. Thus, a π pulse completely inverts the atomic probability, taking the atom from the ground state to the excited state. A 2π *pulse* is one that returns the atom via a 360° rotation of its Bloch vector to its initial state, after passing through the excited state. The remarkable nature of this rotation is not so much that an inversion of the atomic state is possible, but that it can be done fully coherently and without regard for the pulse shape. Only the total integral of $\mathscr{E}_0(t)$ is significant.

The physics behind Bloch vector rotation is essentially the same in magnetic resonance, but perhaps less remarkable since the Bloch vector in that case is a physically "real" magnetic moment. In optical resonance there is no "real" electric moment vector whose Cartesian components can be identified with $[u, v, w]$. Early evidence of the response of the optical Bloch vector to coherent pulses was obtained in experiments of Tang, Gibbs, Slusher, Brewer, and others (see Sections II.B and II.C), and further experiments probing these properties continue to be of interest.

B. Photon Echoes

Photon echoes are an example of spontaneous recovery of a physical property that has been dephased after many relaxation times have elapsed. In the case of echoes the physical property is the macroscopic polarization of a sample of two-level atoms. Recovery of the polarization means recovery of the ability to emit radiation, and the signature of a photon echo is the appearance of a burst of radiation from a long-quiescent sample of atoms. The burst occurs at a precisely predictable time, not randomly, and is due to a hidden long-term memory. The echo principle was discovered and spin echoes were observed by Hahn in 1950 in magnetic resonance experiments. Photon echoes were first observed by Hartmann and co-workers in 1965.

Photon echoes are possible when a sample of atoms is characterized by a broad distribution $g(\Delta)$ of detunings. This may occur in a gas, for example, because of Doppler broadening or in a solid because of crystalline inhomogeneities. For the latter reason it is said that $g(\Delta)$ indicates the presence of *inhomogeneous broadening*. The Maxwellian distribution of velocities in a gas is equivalent to a Maxwellian distribution of detunings since each atom's Doppler shift is proportional to its velocity. If the number of atoms in the sample is N, then the fraction with detuning Δ is given by $Ng(\Delta)\,d\Delta$, where

$$g(\Delta) = \frac{1}{\sqrt{(2\pi)}\delta\omega_{\mathrm{D}}} e^{(-1/2)[(\Delta-\bar{\Delta})^2/(\delta\omega_{\mathrm{D}})^2]}. \tag{26}$$

Here $\bar{\Delta}$ is the average detuning and $\delta\omega_{\mathrm{D}}$ is the Doppler linewidth, $\delta\omega_{\mathrm{D}} = \omega_{\mathrm{L}}(kT/mc^2)^{1/2}$.

The Bloch vector picture is well suited for describing photon echoes. Assume that the Bloch vectors for a collection of N atoms all lie in the equatorial $(u - v)$ plane of the unit sphere along the negative v axis, that is, $\mathbf{S} = [0, -1, 0]$. This arrangement can be accomplished by excitation from the ground state $[0, 0, -1]$ with a strong $\pi/2$ pulse for which $\mathbf{Q} = [-\chi, 0, \Delta] \approx [-\chi, 0, 0]$ if $\chi \gg \Delta$. The total Bloch vector is then $\mathbf{S}_N = [U, V, W] = [0, -N, 0]$, which corresponds to a macroscopic dipole moment of magnitude Nd. After this excitation pulse the sample begins to radiate coherently at a rate appropriate to the dipole moment Nd. However, following the excitation pulse we again have $\chi = 0$ and $\mathbf{Q} = [0, 0, \Delta]$, and according to Eq. (23) the individual Bloch vectors immediately begin to process freely about the w axis (in the $u - v$ plane) at rates depending on their individual detunings Δ. Specifically, $(u - iv)_t = (u - iv)_0 e^{-i\Delta t}$ for an atom with detuning Δ, if $\chi = 0$.

As a consequence, the total Bloch vector will rapidly shrink to zero in a time $\delta t \approx 1/\delta\omega$, where $\delta\omega$ is the spread in angular velocities in the N atom collection. That is, the coherent sum of N dipoles rapidly dephases and $[0, N, 0] \rightarrow [0, 0, 0]$, with the result that the sample quickly stops radiating. This is called *free precession decay*, or free induction decay after the similar effect in magnetic resonance, because the decay is due only to the fact that the individual dipole components $u - iv$ get out of phase with each other due to their different precession speeds, not because any individual dipole moment is decaying.

If the distribution of Δ's is determined by the Doppler effect, as in Eq. (26), then since $(u - iv)_0 = i$, one finds

$$\begin{aligned} U - iV &= iN \int d\Delta\, g(\Delta) e^{-i\Delta t} \\ &= iN e^{-i\bar{\Delta}t} \exp\left[-\tfrac{1}{2}\left(\delta\omega_{\mathrm{D}}^2\right)^2 t\right]. \end{aligned} \tag{27}$$

The decay is very rapid if the Doppler width is large. Typically $\delta\omega_{\mathrm{D}} \approx 10^9$ to 10^{10} s^{-1}, thus, within a few tenths of a nanosecond $U - iV \rightarrow 0$ and radiation ceases.

The echo method consists of applying a second pulse to the collection of atoms after U and V have vanished, that is, at some time $T \gg (\delta\omega_{\mathrm{D}})^{-1}$. Each single atom with its detuning Δ, after the time interval T, still has $(u - iv)_T = ie^{-i\Delta T}$. Only the *sum* of these u and v values is zero due to their different Δ values. The torque vector describing the second pulse is $\mathbf{Q} = [-\chi, 0, \Delta]$, which can again be approximated by $[-\chi, 0, 0]$ if $\chi \gg \Delta$. The effect of the second pulse is again to rotate the Bloch vectors about the u axis. The ideal second pulse is a π pulse, in which case $u \rightarrow u$, $v \rightarrow -v$ and $w \rightarrow -w$, that is, a rotation by 180°. Thus, for times $t \geq T$ after the π pulse the $u - v$ components are

$$(u - iv)_t = (u - iv)_{T'} e^{-i\Delta(t-T)},$$

where T' signifies the rotated coherence vector immediately following the π pulse at time T:

$$(u - iv)_{T'} = (u + iv)_T = -ie^{i\Delta T}.$$

The remarkable feature of a π pulse is that it accomplishes an effective reversal of time. Following it the Bloch vectors do not continue to dephase, but begin to rephase:

$$\begin{aligned} (u - iv)_t &= -ie^{i\Delta T} e^{-i\Delta(t-T)} \\ &= -ie^{-i\Delta(t-2T)}. \end{aligned} \tag{28}$$

Thus, at the exact time $t = 2T$, the individual u's and v's all rephase perfectly: $[u, v, w] = [0, 1, 0]$. Their Bloch vectors are merely rotated 180° from their positions after the original $\pi/2$ pulse, and they again constitute a macroscopic dipole moment $\mathbf{S}_N = [0, N, 0]$, and therefore the collection will begin to radiate again.

Because of the timing of this radiation burst, exactly as long (T) after the π pulse as the π pulse was after the original $\pi/2$ pulse, it is natural to call the signal a *photon echo*. Because of the separation by the intervals T and $2T$ from the π and $\pi/2$ excitation pulses, the observation of an echo can be in practice an observation that is very noise free. Following the echo pulse, the coherence vectors again immediately begin to dephase, but they can again be rephased using the same method, and a sequence of echoes can be arranged.

Because of collisions with other atoms, the individual u and v values will actually get smaller during the course of an echo experiment, independent of their Δ values. Thus, the rephased Bloch vector is not quite as large as the original one. One of the possible uses of an echo experiment is to measure the rate of collisional decay of u and v, say as a function of gas pressure, by measuring the echo intensity in a sequence of experiments with different values of T since the echo intensities will get smaller as collisions reduce the length of the rephased Bloch vector. The way in which Eq. (23) is rewritten to account for collisions is taken up in Section II.D.

C. Self-Induced Transparency and Short Pulse Propagation

The polarization of a dielectric medium of two-level atoms, such as any atomic vapor excited near to resonance, is linearly related to the incident electric field strength $\mathcal{E}_0$ at low-light intensities but becomes nonlinear in the strong-field, short-pulse regime. This is most evident in Eq. (18), where the sine and cosine functions contain all powers of $\chi = 2d\mathcal{E}_0/\hbar$. These nonlinearities have striking consequences for optical pulse propagation.

If a 2π pulse is injected into a collection of on-resonant two-level atoms, it cannot give any energy to them because after its passage the atoms have been dynamically forced back into their initial state. Thus, a 2π pulse has a certain energy stability and so does a 4π pulse and every $2n\pi$ pulse for the same reason. However, all other pulses are obviously not stable since they must give up some energy to the atoms if they do not rotate the atomic chohérence vectors all the way back to their initial positions.

The effect on the injected pulse due to atomic absorptions is given by the Maxwell equation (21). When there is a broad distribution $g(\Delta)$ of detunings among the atoms, then Eq. (21) leads to a so-called area theorem, a nonlinear propagation equation for a pulse of area ϕ:

$$d\phi(z)/dz = -\tfrac{1}{2}N\sigma \sin\phi(z). \tag{29}$$

Here $\phi(z)$ means the total pulse area $\int \chi(z, t')\, dt'$, where the integral extends over the duration of the pulse. The solution is $\tan \phi(z)/2 = e^{-\alpha z/2}$, and the attenuation coefficient is $\alpha = N\sigma$, where N is the density of atoms and σ is the inhomogeneous absorption cross section:

$$\sigma = \int g(\Delta')\sigma_{\mathrm{a}}(\Delta')\, d\Delta', \tag{30}$$

where $\sigma_{\mathrm{a}}(\Delta')$ is the single-atom cross section (see Section II.E). For very weak pusles with $\phi \ll \pi$, one can replace $\sin \phi(z)$ by $\phi(z)$ and recover from Eq. (29) the usual linear law for pulse propagation: $d\phi(z)/dz = -(\alpha/2)\phi(z)$, which predicts exponential attenuation of the pulse: $\phi(z) = \phi(0)\exp[-\alpha z/2]$. The factor of $\frac{1}{2}$ arises because ϕ is proportional to the electric field amplitude, not the intensity.

Remarkably, one of the "magic" pulses with $\phi = 2\pi n$, which does not lose energy while propagating, also preserves its shape. This is the 2π pulse, for which there is a constant-shape solution of the reduced Maxwell–Bloch equations:

$$\chi(z, t) = (2/\tau_p)\mathrm{sech}[(t - z/V)/\tau_p]. \tag{31}$$

That is, the entire pulse moves at the constant velocity V, which can be several orders of magnitude slower than the normal light velocity in the medium. In ordinary light propagation this would correspond to an index of refraction $n \approx 1000$. All of these remarkable features were discovered by McCall and Hahn in the 1960s and labeled by the term *self-induced transparency* to indicate that a light pulse could manipulate the atoms in a dielectric in such a way that the atoms cannot absorb any of the light.

Self-induced transparency is an example of soliton behavior. The nonlinearity of the coupled Maxwell–Bloch equations opposes the dispersive character of normal light transmission in a polarizable medium to permit a steady nondispersing solitary wave (or soliton) [Eq. (31)] to propagate unchanged. This happens only for fields sufficiently strong that the 2π pusle condition can be met. The Maxwell–Bloch equations can in many cases be shown to be equivalent to the sine–Gordon soliton equation or generalizations of it.

D. Relaxation

Both photon echoes and self-induced transparency demonstrate the existence of optical phenomena depending on $\chi \approx \mathcal{E}$, and not on $\mathcal{E}^2$, that is, on the Maxwell–Bloch equations and not on the Einstein rate equations. How are these two approaches to light–matter interactions connected? To answer this question it is necessary to extend the scope of the Bloch equations and include the effects of line-broadening and relaxation processes.

The upper levels of any system have a finite lifetime, and so $|a_2|^2$ cannot oscillate indefinitely as Eq. (14b) implies, but must relax to zero. This is most fundamentally due to the possibility of spontaneous emission of a photon, accompanied by a transition in the system to the lower level. Such transitions occur at the rate A, as in Einstein's equation (1b).

Other relaxation processes also occur. For example, collisions with other atoms cause unpredictable changes in the state of a given two-level atom. These collisional changes typically affect the dipole coherence of the two-level atom instead of the level probabilities, that is, they affect u and v instead of w. We suppose that the rate of such processes is γ. Although γ does not appear in Einstein's rate equations, its existence is implied. This will be clarified later.

The fundamental equations of optical resonance, Eqs. (19), can be rewritten to include these relaxations as follows:

$$du/dt = -\Delta v - (\gamma + A/2)u \tag{32a}$$

$$dv/dt = +\Delta u + \chi w - (\gamma + A/2)v \tag{32b}$$

$$dw/dt = -\chi v - A(w + 1). \tag{32c}$$

In the absence of relaxation ($\gamma = A = 0$), the solutions of Eqs. (32) are purely oscillatory [recall Eq. (18)] and are said to be *coherent*. In the absence of the radiation field ($\chi = 0$), the on-resonance solutions are completely nonoscillatory:

$$u = u_0 e^{-(\gamma + A/2)t}$$
$$v = v_0 e^{-(\gamma + A/2)t}$$
$$w = -1 + (w_0 + 1)e^{-At}, \quad \text{all for } \chi = 0.$$

These solutions are said to be *incoherent*. In each case "coherence" refers to the existence of oscillations with a well-defined period and phase. In Bloch's notation, the relaxation rates are written $\gamma + A/2 = 1/T_2$, where $A = 1/T_1$, and T_1 and T_2 are called the "longitudinal" and "transverse" rates of relaxation.

Relaxation theory is a part of statistical physics, and in quantum theory statistical properties of atoms and fields are usually discussed with the aid of the quantum mechanical density matrix ρ. The density matrix for a two-level atom has four elements, $\rho_{11}, \rho_{12}, \rho_{21}$, and ρ_{22}. These are related to u, v, and w by the equations $u = \rho_{12} + \rho_{21}$, $v = -i(\rho_{12} - \rho_{21})$, and $w = \rho_{22} - \rho_{11}$, which have the inverse forms:

$$\rho_{12} = \tfrac{1}{2}(u + iv) = \langle a_1 a_2^* \rangle \tag{33a}$$
$$\rho_{21} = \tfrac{1}{2}(u - iv) = \langle a_2 a_1^* \rangle \tag{33b}$$
$$\rho_{11} = \tfrac{1}{2}(1 - w) = \langle a_1 a_1^* \rangle \tag{33c}$$
$$\rho_{22} = \tfrac{1}{2}(1 + w) = \langle a_2 a_2^* \rangle. \tag{33d}$$

Here the brackets $\langle \ldots \rangle$ are understood to refer to an average over an ensemble of parameters and variables inaccessible to direct and deterministic evaluation, such as the initial positions and velocities of all the atoms in a collection that may collide with and disturb a typical two-level atom.

The equations given in Eqs. (19) for u, v, and w can also be obtained from ρ via the Liouville equation of quantum statistical mechanics: $i\hbar\, d\rho/dt = [H, \rho]$. The equations for the density matrix elements [Eqs. (33)] are

$$\begin{aligned} d\rho_{21}/dt &= -(\gamma + A/2 + i\Delta)\rho_{21} \\ &\qquad - (i\chi/2)(\rho_{22} - \rho_{11}) \end{aligned} \tag{34a}$$
$$\begin{aligned} d\rho_{12}/dt &= -(\gamma + A/2 - i\Delta)\rho_{12} \\ &\qquad + (i\chi/2)(\rho_{22} - \rho_{11}) \end{aligned} \tag{34b}$$
$$d\rho_{11}/dt = A\rho_{22} - (i\chi/2)(\rho_{12} - \rho_{21}) \tag{34c}$$
$$d\rho_{22}/dt = -A\rho_{22} + (i\chi/2)(\rho_{12} - \rho_{21}). \tag{34d}$$

We now demonstrate the connection between Einstein's equations and the quantum optical Eqs. (34). Consider the weak-field limit $\chi \ll |\gamma + A/2 + i\Delta|$. In this limit the rate of change of the "off-diagonal" density matrix elements ρ_{21} and ρ_{12} is dominated by the first factor $-(\gamma + A/2 \pm i\Delta)$, and both ρ_{21} and ρ_{12} decay rapidly. If in addition $A \ll |\gamma + A/2 \pm i\Delta|$, then ρ_{22} and ρ_{11} change relatively slowly, and so ρ_{21} and ρ_{12} rapidly adjust themselves to the small quasi-steady values:

$$\rho_{21} \approx -(i/2)\chi[\gamma + A/2 + i\Delta]^{-1}(\rho_{22} - \rho_{11}) \tag{35a}$$
$$\rho_{12} \approx (i/2)\chi[\gamma + A/2 - i\Delta]^{-1}(\rho_{22} - \rho_{11}). \tag{35b}$$

These solutions show that the off-diagonal elements of the density matrix can be determined from constant numerical factors and combinations of the diagonal density matrix elements. They can then be eliminated from Eqs. (34c and d).

This procedure is referred to as *adiabatic elimination* of off-diagonal coherence because the remaining equations for ρ_{11} and ρ_{22} no longer exhibit coherence. That is, they no longer have oscillatory solutions. The term *adiabatic* is appropriate in the sense that ρ_{21} and ρ_{12} are entrained by the slower ρ_{11} and ρ_{22}. The reverse procedure, the elimination of ρ_{11} and ρ_{22} in favor of ρ_{21} and ρ_{12}, is not possible because the reverse inequality $|\gamma + A/2 \pm i\Delta| \ll A$ is not possible.

These adiabatic off-diagonal solutions, once inserted into Eqs. (34) lead to an equation for the slowly changing ρ_{22} as follows:

$$\frac{d\rho_{22}}{dt} = -A\rho_{22} - \left[\frac{1}{2}\chi^2 \frac{\gamma + A/2}{\Delta^2 + (\gamma + A/2)^2}\right](\rho_{22} - \rho_{11}). \tag{36}$$

Recall $\rho_{22} = |a_2|^2$ is the probability that the atom is in its upper state, and thus plays the same role as P_n in the Einstein equation (2a). Similarly ρ_{11} plays the role here of P_m there. By comparing Eqs. (2a) and (36) one sees that they are identical in form and content if one identifies the coefficient of ρ_{11} in Eq. (36) with the coefficient of P_m in Eq. (2a). In other words, the density matrix equations [Eqs. (34a–d)] of quantum optics contain Einstein's equation in the weak-field and adiabatic limits $\chi \ll |\gamma + A/2 \pm \Delta|$ and $A \ll |\gamma + A/2 \pm i\Delta|$. The B coefficient can be derived in this limit (see Section II.E) if one properly interprets the factor $\frac{1}{2}\chi^2(\gamma + A/2)/[\Delta^2 + (\gamma + A/2)^2]$.

Relaxation processes affect the Maxwell field as well as the Bloch variables. To determine the form of relaxation to assign to Maxwell equation (21) it is sufficient to consider the conservation of energy. From Eqs. (21) and (19c) it follows that

$$[\partial/\partial z + \partial/\partial ct]|\mathscr{E}|^2 = -(N\hbar\omega/4\varepsilon_0 c)[\partial\omega/\partial t],$$

which is equivalent to an equation for photon flux and level probability

$$[\partial/\partial z + \partial/\partial ct]\Phi(z, t) = -N\partial P_2/\partial t, \tag{37a}$$

since $\hbar\omega\Phi \equiv I = 2c\varepsilon_0|\mathscr{E}|^2$ and $w = 2P_2 - 1$.

Equation (37a) is Poynting's theorem for a "one-dimensional" medium. It expresses the conservation of photon flux in terms of atomic excitations. However, in the absence of the resonant two-level atoms ($N = 0$), Eqs. (37) predicts $\Phi(z, t) = \Phi_0(z - ct)$, which means that the photon flux has the constant value $\Phi(z_0)$ at every point $z = z_0 + ct$ that travels with the pulse. This is contradictory to ordinary experience in two respects. The medium that is host to the two-level atoms (e.g., other gas atoms, a solvent, or a crystal lattice) always causes both dispersion and absorption. They can be taken into account by modifying Eq. (37) slightly:

$$[\partial/\partial z + \kappa + \partial/\partial v_g t]\Phi(z, t) = -N\partial P_2/\partial t, \qquad (37b)$$

where v_g is the group velocity for light pulses in the medium and κ its linear attenuation coefficient.

This form of the flux equation implies a similar alteration of Maxwell's equation (21):

$$[\partial/\partial z + \kappa/2 + \partial/\partial v_g t]\mathscr{E} = i(Nd\omega/4\varepsilon_0 c)[u - iv]. \qquad (38)$$

This form of Maxwell's equation is useful in describing the elements of laser theory (see Section II.G).

E. Cross Section and the *B* Coefficient

How does one use quantum optical expressions to obtain basic spectroscopic formulas, such as for the absorption cross section and B coefficient? That is, given expressions derived from Eqs. (34), which are based on the Rabi frequency χ instead of the more familiar radiation intensity I or spectral energy density $u(\omega)$, how does one recover a cross section, for example? Consider the quantum optical derivation of the Einstein formula in Eq. (36). The transition rate (absorption rate or stimulated emission rate) can be identified readily. With the abbreviation $\beta = \gamma + A/2$, one obtains:

$$\text{abs. rate} = \frac{1}{2}\chi^2 \frac{\beta}{\Delta^2 + \beta^2}. \qquad (39a)$$

The absorption rate is a peaked function of Δ whose value drops to $\frac{1}{2}$ of the maximum value at $\Delta = {}^{\pm}\beta$. Thus, β is called the *halfwidth at halfmaximum* (HWHM) of the absorption lineshape. This shows that relaxation leads to *line broadening*, and since β applies equally and individually to every atom, it is an example of a *homogeneous linewidth*. Recall (Section I.B) that inhomogeneous broadening is not a characteristic of individual atoms but of a collection of them. From expression (39a) at exact resonance one obtains the relationship:

$$\text{resonant transition rate} = \chi^2/\text{full linewidth}.$$

This is the single most concise relationship between the parameters of incoherent optical physics (transition rate and linewidth) and the central parameter of coherence (Rabi frequency). It holds in situations much more general than the present example and allows rapid and accurate translation of formulas from one domain to the other.

With the use of $\chi = 2d\mathscr{E}/\hbar$ and $I = 2c\varepsilon_0\mathscr{E}^2$, Eq. (39a) becomes

$$\text{abs. rate} = \frac{D^2 I}{3\varepsilon_0 c\hbar^2}\frac{\beta}{\Delta^2 + \beta^2}. \qquad (39b)$$

The introduction of the new dipole parameter D here is based on the assumption that all orientations of the atomic dipole matrix element $\mathbf{d}_{21}$ are possible (in case, e.g., all magnetic sublevels of the main levels 1 and 2 are degenerate). Then $d^2 \equiv |\mathbf{e}\cdot\mathbf{d}_{21}|^2$ must be replaced by its spherical average, that is, by $D^2/3$, where $D^2 = |\mathbf{d}_{21}\cdot\mathbf{d}_{12}|$. We assume this is appropriate in the remainder of Section II.

The atomic absorption cross section $\sigma_{\rm a}$ is, by definition, the ratio of the rate of energy absorption, $\hbar\omega_{21} \times$ (abs. rate), to the energy flux (intensity) I of the photons being absorbed. From Eq. (39b) this ratio is

$$\sigma_{\rm a}(\omega; \omega_{21}) = \frac{D^2\omega}{3\varepsilon_0\hbar c}\frac{\beta}{(\omega_{21} - \omega)^2 + \beta^2}, \qquad (40)$$

where $\gamma = \beta - A/2$ is the specifically collisional contribution to the halfwidth. Formula (40) shows that a quantum optical approach to light absorption through the weak field limit of Eqs. (34) gives conventional results of atomic spectroscopy. If Doppler broadening is present, then Eq. (40) must be integrated over the Doppler distribution of detunings Eq. (26) as was done in Eq. (30).

Lineshape plays a key role in understanding the relationship of Eqs. (39) to the Einstein expression (1a) relating absorption rate to the B coefficient. The absorption cross section can be written $\sigma_{\rm a} = \sigma_{\rm t} S_{\rm a}(\omega; \omega_{21})$, where $\sigma_{\rm t}$ is the total frequency-integrated cross section:

$$\sigma_{\rm t} = \pi D^2\omega_{21}\big/3\varepsilon_0\hbar c \qquad (41a)$$

and $S_{\rm a}$ is the atomic lineshape

$$S_{\rm a}(\omega; \omega_{21}) = \frac{\beta/\pi}{(\omega_{21} - \omega)^2 + \beta^2}, \qquad (41b)$$

which is normalized according to $\int d\omega\, S_{\rm a} = 1$, and in this case has a Lorentzian shape. A lineshape also exists for the radiation field and is expressed by $u(\omega)$, the spectral energy density function. One connects $u(\omega)$ with I by the frequency integral $c\int u(\omega')\,d\omega' = I$. In the monochromatic case $cu(\omega')$ takes the idealized singular form $cu(\omega') = I\delta(\omega - \omega')$, and Eq. (39b) is the result for the absorption rate.

In the general nonmonochromatic case, the expression for absorption rate involves the integrated overlap of $u(\omega)$ and the atomic lineshape function:

$$\text{abs. rate} = \frac{c\sigma_1}{\hbar\omega_{21}}\int S_{\rm a}(\omega'; \omega_{21})u(\omega')\,d\omega'. \qquad (42)$$

This has the desired limiting form, involving $u(\omega)$, and not $I = c\int u(\omega)\,d\omega$, if the spectral width $\delta\omega_L$ of $u(\omega)$ is very broad in comparison to the width β of the absorption lineshape S_a. In this limit, which is implicit in Einstein's discussion, S_a acts in Eq. (42) like a δ-function peaked at $\omega' = \omega_{21}$, and $u(\omega)$ is evaluated at $\omega = \omega_{21}$. From Eq. (42) one can then extract the Einstein B coefficient:

$$B = \frac{\pi D^2}{3\varepsilon_0 \hbar^2}. \tag{43}$$

A simple relation obviously exists between B and σ_t, namely $\hbar\omega_{21} B = c\sigma_t$.

Another quantity of interest is $\sigma_a(0)$, the on-resonance or peak cross section:

$$\sigma_a(0) = \frac{D^2\omega_{21}}{3\varepsilon_0 \hbar c\beta}. \tag{44}$$

By definition, $\sigma_a(0) = \sigma_a(\omega = \omega_{21})$; or, conversely, $\sigma_a(\omega;\, \omega_{21}) = \pi\beta\sigma_a(0)S_a(\omega;\, \omega_{21})$. Representative values of $\sigma_a(0)$ for an optical resonance transition lie in the range $\sigma_a(0) \approx 10^{-13}$ to 10^{-17} cm^2 for absorption linewidths in the range $\beta \approx 10^8$ to 10^{11} s^{-1}.

F. Strong Field Criterion and Saturation

The inequality $\chi \ll |\gamma + A/2 \pm i\Delta|$, on which the absorption rate formula (39) and thus the Einstein B coefficient is based, is important in quantum optics and radiation physics generally because it provides a criterion for distinguishing weak radiation fields from strong radiation fields. The inequality implies that there is a critical value $\mathscr{E}_{cr}$ for field strength that gives a universal meaning to the terms *weak field* and *strong field*, namely $\mathscr{E}_0 \ll \mathscr{E}_{cr}$ and $\mathscr{E}_0 \gg \mathscr{E}_{cr}$, respectively, where:

$$\begin{aligned}\mathscr{E}_{cr} &= (\hbar/d)|\gamma + A/2 \pm i\Delta| \\ &= (\hbar/d)|\beta \pm i\Delta|. \end{aligned}\tag{45}$$

However, since the parameters γ, A, d, and Δ may vary by many orders of magnitude from case to case, the numerical value of $\mathscr{E}_{cr}$ may fall anywhere in an extremely wide range. Thus, it is possible that in one experiment a laser with the power level 10^{20} W/m^2 must be designated "weak," while another laser in a different experiment with the power level 1 W/m^2 must be considered "strong." This factor of 10^{20} is one indication of the great extent of the domain of quantum optics.

In conventional spectroscopy one sometimes encounters saturation effects. These are of course strongest in the strong field regime and are of interest in quantum optics.

There are two distinct time regimes of saturation phenomena. If $\chi \gg \beta$, then there is a range of times $t \ll \delta T \;\approx\; \beta^{-1}$ that can still contain many Rabi oscillations since $\delta T \gg \chi^{-1}$. During the time $0 \leqq t \ll \delta T$, the fully coherent undamped formula (14b) can be used for the upper state probability. On average, during this time the probability that the atom is in its upper level is

$$P_2 = \tfrac{1}{2}\chi^2/[\Delta^2 + \chi^2] \quad \text{(short time average)}, \tag{46a}$$

which is a Lorentzian function of $\Delta = \omega_{21} - \omega$ with the power-broadening halfwidth $\delta\omega_p = \chi$. Power broadening is a saturation effect, because if $\chi \gg \Delta$, then $P_2 \to \frac{1}{2}$ on average, which is obviously saturated, that is, unchanged if χ is made still larger.

Another saturation regime exists for long times, $t \gg \delta T \approx \beta^{-1}$. The solution of Eqs. (24) for $P_2 = \rho_{22}$ in this limit is

$$P_2 = \frac{1}{2}\chi^2 \frac{\beta/A}{\Delta^2 + \beta^2 + \chi^2\beta/A}. \tag{46b}$$

In this case χ begins to dominate the width then $\chi^2 > \beta A$, which defines the saturation value of χ:

$$\chi_{\text{sat}} = \sqrt{(\beta A)}. \tag{47}$$

The power-broadening part of the width of Eq. (46b) is different than in Eq. (46a), namely $\delta\omega_p = \chi\sqrt{(\beta/A)}$. Depending on the value of $\gamma \equiv \beta - A/2$, the power width here can be anything between a minimum of $\chi/\sqrt{2}$, if $\gamma = 0$, and a maximum of $\chi\sqrt{(\gamma/A)}$, if $\gamma \gg A$. This distinction between the saturated power-broadened linewidth $\delta\omega_p$ predicted by Eq. (46a) for short times and by Eq. (46b) for asymptotically long times has caused some confusion in the past. Further study indicates that Eq. (46b) breaks down for sufficiently large χ, basically because Bloch-type relaxation, such as assumed in Eqs. (32), becomes invalid. The first experimental reports of this regime of optical resonance were made by Brewer, De Voe, Mossberg, and others in the early 1980s.

In the case of asymptotically long times the expression for P_2 can be written in several ways, using Eqs. (47) or (40):

$$\begin{aligned} P_2 &= \frac{\frac{1}{2}(\chi/\chi_{\text{sat}})^2}{1 + (\Delta/\chi_{\text{sat}})^2 + (\chi/\chi_{\text{sat}})^2} \\ &= \frac{\sigma_a I}{\hbar\omega_{21}A + 2\sigma_a I} \\ &= \frac{\Phi/\Phi_{\text{sat}}}{1 + 2\Phi/\Phi_{\text{sat}}}, \end{aligned}\tag{48}$$

where $\Phi = I/\hbar\omega_{21}$ and we have introduced the saturation flux required for saturation of the transition $\Phi_{\text{sat}} = A/\sigma_a$ (or $= 1/T_1\sigma_a$ in Bloch's notation, which is more appropriate if there are other contributions than spontaneous emission rate A to the level lifetimes). Figure 3 shows the effect of both power broadening and saturation on the steady-state probability P_2.

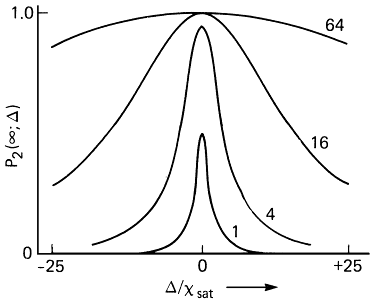


FIGURE 3 Population saturation for different values of χ/χ_{sat}.

In common with all two-level saturation formulas, Eq. (48) and Fig. 3 predict $P_2 = \frac{1}{2}$ at most. However, this prediction is valid only for weak fields or for long times. As the solutions in Eq. (14) and in Fig. 2 show, strong monochromatic resonance radiation can repeatedly transfer the electron to the upper level with $P_2 \approx 1$ for times $\ll \beta^{-1}$. Experiments that show $P_2 > \frac{1}{2}$ have been practical only with lasers. Laser pulses are both short and intense, allowing $\chi \gg \beta$ as well as $t \ll \beta^{-1}$.

G. Semiclassical Laser Theory

The coupled Maxwell–Bloch equations can be used as the basis for laser theory. It is a semiclassical theory, but still adequate to illustrate the most important results, such as the roles of inversion and feedback, the existence of threshold, and the presence of frequency pulling in steady-state operation. A fully detailed semiclassical theory of laser operation was already given by Lamb in 1964.

The density matrix equations (24) must be modified to allow for pumping of the upper laser level and to allow the lower level to decay to a still lower level labeled 0 and not previously needed. The rates of these new processes are denoted R and Γ, respectively. This is basically the scheme of a so-called three-level laser, as sketched in Fig. 4.

The diagonal equations change to

$$d\rho_{11}/dt = -\Gamma\rho_{11} + A\rho_{22} + (i/2)\chi_m(\rho_{12} - \rho_{21}) \quad (49)$$

$$d\rho_{22}/dt = -A\rho_{22} - (i/2)\chi_m(\rho_{12} - \rho_{21}) + R. \quad (50)$$

Here the index m or χ_m shows that we are dealing with the electric field of the mth mode of the laser cavity. The laser is easy to operate only if $\Gamma \gg A$ because only then can a large positive inversion be maintained between the two levels with a value of R that is not too high and without seriously depleting the population of level 0. Under other conditions, the equation for the density matrix element ρ_{00} would also have to be included. The off-diagonal density matrix equations (24a and b) are basically unchanged if we interpret γ as including another contribution $\Gamma/2$, where Γ is the decay rate of the lower level probability to level 0 in Fig. 4.

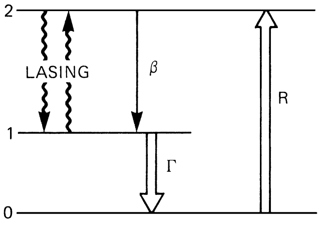


FIGURE 4 Schematic diagram of three-level laser system.

The reduced Maxwell equation (28) is now useful. We will ignore the difference between v_g and c, and write Eq. (38) in terms of χ instead of $\mathcal{E}$, and use $\rho_{21} = \frac{1}{2}(u - iv)$ to find

$$[\partial/\partial z + \kappa/2 + \partial/\partial ct]\chi_m = i\big(Nd^2\omega/\varepsilon_0\hbar c\big)\rho_{21}. \quad (51)$$

Note that in a cavity κ can arise principally from mirror losses and not from absorption in the cavity volume. The same adiabatic elimination of dipole coherence undertaken in Eqs. (35) provides the value for ρ_{21} to insert into Eq. (51), which in steady state ($\partial\chi_m/\partial t = 0$) becomes:

$$\big[\partial/\partial z + \kappa/2 - \tfrac{1}{2}(g + i\delta\kappa)\big]\chi_m = 0, \quad (52)$$

where

$$g + i\delta\kappa = \big(ND^2\omega/3\varepsilon_0\hbar c\big)[w_{\text{ss}}/(\beta + i\Delta)] \quad (53)$$

and now $\beta = \gamma + (A + \Gamma)/2$. By using Eq. (40) one can obtain

$$g = N\sigma_{\text{a}}(\Delta)w_{\text{ss}} \quad (54a)$$

$$c\delta\kappa = -(gc/\beta)(\omega_{21} - \omega), \quad (54b)$$

where Eq. (54a) has been used to simplify Eq. (54b). Here w_{ss} denotes the inversion $\rho_{22} - \rho_{11}$ in steady state.

It is clear that g is the intensity gain coefficient, since if $g > \kappa$, Eq. (52) would predict exponential growth, $|\chi_m|^2 \approx \exp[(g - \kappa)z]$, in an open-ended medium such as a laser amplifier. It is thus clear that $g = \kappa$ is the threshold condition for amplification or laser operation. Also obviously, g is not positive unless w_{ss} is positive. Recall that in an ordinary noninverted medium $w = -1$, and in this case $g = -N\sigma_{\text{a}} = -\alpha$, where α is the ordinary absorption coefficient.

Rather than growing indefinitely, the field in a laser cavity must conform to the spatial period determined by the mirrors. Thus, at steady state $\chi_m(z) \approx \chi_0 \exp[i\Delta k_m z]$, where the phase $\Delta k_m z$ is the difference between the actual phase of steady-state laser operation $k_m z$ and the phase $kz = \omega z/c$ that was assumed initially in defining the field carrier wave and envelope functions in Eq. (20). Since χ_m does not depend on the transverse coordinates x and y, this theory cannot describe transverse mode structure, and $k_m = m\pi/L$, where L is the cavity length and m is the longitudinal mode number. Operating values could be $m \approx 10^6$ and $L \approx 10$ cm, in which case the laser would run at a frequency near to $m\pi c/L \approx 3 \times 10^{15}\ \mathrm{s}^{-1} \approx 5 \times 10^{14}$ Hz.

In laser operation $\partial/\partial z$ can be replaced in Eq. (52) by $i\Delta k_m$ and the imaginary part of Eq. (52) becomes $\Delta k_m = \frac{1}{2}\delta\kappa$, which leads directly to a condition for the operating frequency ω:

$$\omega_m - \omega = -(\kappa c/2\beta)(\omega_{21} - \omega). \tag{55}$$

This requires ω to lie somewhere between the empty cavity frequency $\omega_m = m\pi c/L$ and the natural transition frequency ω_{21} of the atom, and it is said that the two-level laser medium "pulls" the operating frequency away from the cavity frequency $m\pi c/L$. The solution of Eq. (55) for ω is

$$\omega = \left[\beta\omega_m + \tfrac{1}{2}\kappa c\omega_{21}\right] / \left[\beta + \tfrac{1}{2}\kappa c\right]. \tag{56}$$

H. Optical Bistability

The input–output relationships between light beams injected into and transmitted through an empty optical cavity are linear relationships. However, the situation changes dramatically if the cavity contains atoms. This is obvious in the case of laser action. However, even if the atoms are not pumped, their nonlinearities can be significant.

Consider a laser beam injected into an optical cavity filled with two-level atoms. For simplicity we consider the case where the frequencies of the atoms, the cavity, and the injected laser light are all equal: $\omega = \omega_c = \omega_{21}$. Only the $u - w$ Bloch equation are needed then:

$$dv/dt = -\beta v + (\chi_m + \chi_0)w \tag{57a}$$

$$dw/dt = -A(1 + w) - (\chi_m + \chi_0)v. \tag{57b}$$

Here χ_0 and χ_m are the Rabi frequencies associated with the injected field strength and the cavity mode field generated by the atoms, and $\beta = \gamma + A/2$. The Maxwell equation for the internally generated χ_m is

$$(\kappa/2 + \partial/\partial ct)\chi_m = \left(Nd^2\omega / 2\varepsilon_0\hbar c\right)v. \tag{58}$$

Note that there is no term $i\Delta\kappa_m$ from $\partial/\partial z$, as there is in the discussion of laser operation, only because of the three-way resonance assumption.

The question is, how does the presence of the atoms in the cavity affect the transmitted signal? Since the transmitted signal differs only by a factor of mirror transmissivity from the total field in the cavity $\chi_t = \chi_m + \chi_0$, we ask for the relation between χ_0 and χ_t. The dynamical evolution of the system is complicated. Early attention was given to this situation in the 1970s by Szöke, Bonifacio, Lugiato, McCall, and others.

As with the laser, steady state is sufficiently interesting, so we put $dv/dt = dw/dt = d\chi/dt = 0$ and solve for χ_0 or χ_t. They obey a simple but nonlinear relation:

$$\chi_0 = \chi_t + (\alpha/\kappa)(\beta A)\left[\chi_t / \left(\beta A + \chi_t^2\right)\right], \tag{59}$$

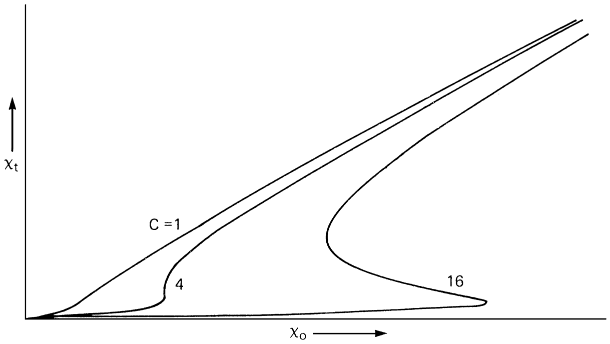


FIGURE 5 Bistable input–output curves for different values of $C = a/2k$.

where α is the (on-resonance) two-level medium's absorption coefficient, $\alpha = ND^2\omega/3\varepsilon_0\hbar c\beta$, and $\beta A = \chi_{\mathrm{sat}}^2$, the saturation parameter identified in Eq. (47). If both χ_0 and χ_t are normalized with respect to χ_{sat} by defining $\xi = \chi_0/\chi_{\mathrm{sat}}$ and $\eta = \chi_t/\chi_{\mathrm{sat}}$ then one finds the dimensionless relation:

$$\xi = \eta + (\alpha/\kappa)[\eta/(1 + \eta^2)]. \tag{60}$$

Of course the inverse relation $\eta = \eta(\xi)$, that is, the total field strength as a function of the input field strength, is more interesting. Figure 5 shows the important features of this relation. It demonstrates that the total field η is double-valued as a function of input field ξ if α/κ is larger than a certain critical value. In this simple model the critical value is $\alpha/\kappa = 8$, and the vertical segment in the central curve is an indication of this. The double-valued nature of the curves for $\alpha/\kappa > 8$ is termed *optical bistability*. In the bistable region of the third curve hysteresis can occur, and a hysteresis loop is shown.

The elements of a primitive optical switch are evident in the bistable behavior shown here. If the input field is held near to the lower turning point, then a very small increase in ξ can lead to a very large jump in η, the transmitted field. The possibility of optical logic circuits and eventually optical computers is clearly suggested even by the simple model described here, and efforts being made around the world to realize these possibilities in a practical way have already achieved limited success.

III. RADIATION COHERENCE AND STATISTICS

A. Coherence of Light

In quantum optics the coherence of light is treated statistically. The need for a statistical description of light, whether quantized or classical, is practically universal since all light beams, even those from well-stabilized lasers, have certain residual random properties that are not uniquely determined by known parameters. These random properties lead to fluctuations of light. A satisfactory description can be based on a scalar electric field:

$$E(\mathbf{r}, t) = E^{(+)}(\mathbf{r}, t) + E^{(-)}(\mathbf{r}, t), \qquad (61a)$$

where $E^{(+)}$ is the *positive frequency part* of E. That is, $E^{(+)}$ is the inverse Fourier transform of the positive frequency half of the Fourier transform of E. From this definition, one has $[E^{(-)}(\mathbf{r}, t)]^* \equiv E^{(+)}(\mathbf{r}, t)$. The split-up into positive and negative frequency parts $E^{(\pm)}(\mathbf{r}, t)$ is motivated by the great significance of quasi-monochromatic fields, for which one can write

$$E^{(+)}(\mathbf{r}, t) = \mathscr{E}(\mathbf{r}, t)e^{-i\omega t} \qquad (61b)$$

$$E^{(-)}(\mathbf{r}, t) = \mathscr{E}(\mathbf{r}, t)^* e^{i\omega t}. \qquad (61c)$$

The term *quasi-monochromatic* means that $\mathscr{E}(\mathbf{r}, t)$ is only slowly time dependent, that is, $|d\mathscr{E}/dt| \ll \omega|\mathscr{E}|$.

The intensity of the light beam associated with this electric field is given by:

$$\begin{aligned} I(\mathbf{r}, t) &= c\varepsilon_0 E(\mathbf{r}, t)^2 \\ &= c\varepsilon_0[2|\mathscr{E}(\mathbf{r}, t)|^2 + \mathscr{E}^2(\mathbf{r}, t)e^{-2i\omega t} + \text{c.c.}], \qquad (62) \end{aligned}$$

where c.c. means complex conjugate. In principle I is rapidly time dependent, but the factors $e^{\pm 2i\omega t}$ oscillate too rapidly to be observed by any realistic detector. That is, a photodetector can respond only to the average of Eq. (62) over a finite interval, say of length T beginning at t, where $T \gg 2\pi/\omega$. The $e^{\pm 2i\omega t}$ terms average to zero and one obtains:

$$\bar{I}_T(t) = (1/T)\int_0^T \boldsymbol{I}(t + t')\,dt' = 2c\varepsilon_0|\mathscr{E}(t)|^2. \qquad (63)$$

The overbar thus denotes a coarse-grained average value that is not sensitive to variations on the scale of a few optical periods. We have dropped the $\mathbf{r}$ dependence as a simplification. If the beam is steady and the averaging interval T is long enough, then both the length T and the beginning value $I(t)$ are unimportant, and $\bar{I}_T(t)$ is also independent of t. This is the property of a class of fields called *stationary*. Obviously even nonstationary fields can be considered stationary in the sense of a long T average, and we will adopt stationarity as a simplification of our discussion and write $\bar{I}_T(t)$ simply as $\bar{I}$.

Consider now the operation of a *Michelson interferometer* (Fig. 6). An incident beam is split into two beams a and b, and after traveling different path lengths, say ℓ and $\ell + \delta$, the beams are recombined and the beam intensity $I_{a+b} = c\varepsilon_0[E_a(t) + E_b(t)]^2$ is measured. If the beam splitter sends equal beams with field strength $E(t)$ and intensity $\bar{I}$ into each path and there is no absorption during the propagation, then one measures

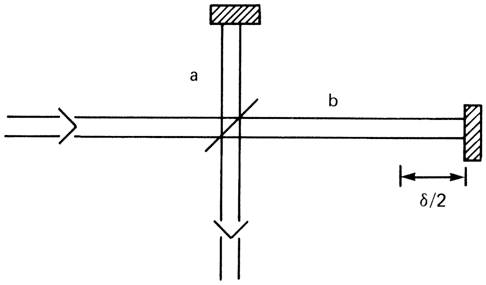


FIGURE 6 Sketch of Michelson interferometer.

$$\begin{aligned} \bar{I}_{a+b} &= (T)^{-1}\int_0^T c\varepsilon_0[E_a(t + t') + E_b(t + t')]^2 dt' \\ &= (T)^{-1}\int_0^T c\varepsilon_0[E(t + t') + E(t + \delta/c + t')]^2 dt' \\ &= \bar{I} + \bar{I} + 4c\varepsilon_0 \\ &\quad \times \text{Re}[\langle\varepsilon^*(t)\varepsilon(t + \delta/c)\rangle e^{-i\omega\delta/c}], \qquad (64) \end{aligned}$$

where Re means "real part" and the angular brackets indicate the time average. If $\delta = 0$, then $\bar{I}_{a+b} = 4\bar{I}$, because the two beams interfere fully constructively. The quantity $\Gamma(\delta/c) = \langle\varepsilon^*(t)\varepsilon(t + \delta/c)\rangle e^{-i\omega\delta/c}$ is called the *mutual coherence function* of the electric field, and the appearance of fringes at the output plane of the interferometer is due to the variation of Γ with δ. From the factor $e^{-i\omega\delta/c}$ it is clear that a fringe shift (a shift from one maximum of $\bar{I}_{a+b}$ to the next) corresponds to a shift of δ by $2\pi c/\omega = \lambda$.

The mutual coherence function is conveniently normalized to its maximum value which occurs when $\delta = 0$, and the normalized function $\gamma(\delta/c)$ is called the *complex degree of coherence*. That is,

$$\gamma(\delta/c) = \langle\varepsilon(t)\varepsilon^*(t + \delta/c)\rangle e^{+i\omega\delta/c}/\langle\varepsilon(t)\varepsilon^*(t)\rangle \qquad (65)$$

and the output intensity can be written:

$$\bar{I}_{a+b} = 2\bar{I}[1 + \text{Re}\,\gamma(\delta/c)], \qquad (66)$$

where Re γ must satisfy $-1 \leqq \text{Re}\,\gamma \leqq 1$.

It is common experience that if the path difference δ is made too great in the interferometer, the fringes are lost and the output intensity is simply the sum of the intensities in the two beams. This can be accounted for by introducing a *coherence time* τ for the light by writing

$$\gamma(\delta/c) = \gamma(0)e^{-|\delta|/c\tau}e^{+i\omega\delta/c}. \qquad (67)$$

This representation for γ has the correct behavior since it vanishes whenever δ is large enough, specifically whenever $\delta \gg c\tau$. For obvious reasons, $c\tau$ is called the *coherence length* of the light. This does not explain the fundamental origin of the coherence time τ, but the lack of such a deep understanding of the light beam can be one reason that a statistical description is necessary in the first place. Typically one adopts Eq. (67) as a convenient empirical relation and interprets τ from it.

The fringe visibility is usually the important quantity that describes an interference pattern, not the absolute level of intensity. The *visibility* is defined by $V = (\bar{I}_{\max} - \bar{I}_{\min})/(\bar{I}_{\max} + \bar{I}_{\min})$, and this is directly related to the complex degree of coherence. Since

$$\operatorname{Re}\gamma = |\gamma(0)|e^{-|\delta|/c\tau}\cos(\omega\delta/c + \phi)$$

one has

$$V = \frac{4\bar{I}|\gamma(0)|e^{-|\delta|/c\tau}}{4\bar{I}} = |\gamma(\delta/c)|. \qquad (68)$$

Thus, the magnitude of the complex degree of coherence is a directly measurable quantity.

One of the foundations of quantum optics was established by Wolf and others in classical coherence theory when it became understood in the 1950s how to describe optical interference effects in terms of measurable autocorrelation functions such as γ. In a sense, the first example of this was provided much earlier by Wiener in 1930 when he showed that the spectrum $S(\omega)$ of a stationary light field is given essentially by the Fourier transform of γ, considered as a function of a time difference τ rather than a path difference δ:

$$S(\omega) = 2\bar{I}\operatorname{Re}\int d\tau' e^{-i\omega\tau'}\gamma(\tau'). \qquad (69)$$

As a specific example, suppose a light beam with carrier frequency ω_L has a Gaussian degree of coherence, with coherence time τ, that is, $\gamma(\tau') = |\gamma(0)|e^{i\omega_L\tau'} \exp[-\frac{1}{2}(\tau'/\tau)^2]$. This is another example, similar to the exponential $\gamma(\delta/c)$ given above, in which a simple analytic function is used to model the normal fact that correlation functions have finite coherence, that is, that $\gamma(\tau) \to 0$ as $\tau \to \infty$. In this example the spectrum is Gaussian:

$$S(\omega) = 2\bar{I}|\gamma(0)|\sqrt{(2\pi\tau^2)}\,\exp\left[-\tfrac{1}{2}(\omega - \omega_L)^2\tau^2\right]. \qquad (70)$$

The effective bandwidth is the frequency range over which $S(\omega)$ is an appreciable fraction of its peak value. In this case the spectrum is centered at $\omega = \omega_L$, and the bandwidth is given by $\Delta\omega = 2\pi\Delta\nu = 1/\tau$.

This is an example of the general rule that the bandwidth is the inverse of the coherence time. A laser beam with bandwidth $\Delta\nu = 100$ MHz has a coherence time $\tau = 1.6$ ns and a coherence length $\Delta\delta = c\tau = 0.5$ m, while sunlight with a bandwidth six orders of magnitude broader ($\Delta\nu = 10^{14}$ Hz) has a coherence time $\tau = 1.6 \times 10^{-3}$ ps $= 1.6$ fs and a coherence length $c\tau = 0.5\,\mu$m that are six orders of magnitude smaller. In the latter case, $\Delta\nu \approx \nu$, and $c\tau \approx \lambda$, so sunlight cannot be called quasi-monochromatic in any sense.

A hierarchy of correlation functions can be defined for statistical fields. One denotes by the *degree of first-order coherence* the normalized first-order correlation of positive and negative frequency parts of the field:

$$g^{(1)}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\left\langle E^{(+)}(\mathbf{x}_1)E^{(-)}(\mathbf{x}_2)\right\rangle}{\left[\left\langle\left|E^{(+)}(\mathbf{x}_1)\right|^2\right\rangle\left\langle\left|E^{(-)}(\mathbf{x}_2)\right|^2\right\rangle\right]^{1/2}}, \qquad (71)$$

where $\mathbf{x} = (\mathbf{r}, t)$ for short. In the case of a quasimonochromatic field $g^{(1)}$ is just the same as γ defined above. The terms "first-order coherent," "partially coherent," and "incoherent" refer to light with $|g^{(1)}|$ equal to 1, between 1 and 0, and 0, respectively.

Definition (71) and other average values can be interpreted in an ensemble sense, as well as in a time average sense. In the ensemble interpretation the angular brackets $\langle \ldots \rangle$ are associated with an average over "realizations," that is, over possible values of the fields at given points $\mathbf{x}$, weighted by appropriate probabilities. Two important examples are (1) so-called chaotic fields, fields that are stationary complex Gaussian random processes, and (2) constant intensity fields. The chaotic distribution is characteristic of ordinary (thermal) light such as omitted by the sun, flames, light bulbs, etc. The constant intensity distribution is an idealization associated with highly stabilized single-mode laser (coherent) light. The ergodic assumption that time and ensemble averages are equal is usually made. We will write $\langle I \rangle$ and $\bar{I}$ interchangeably.

If $p[\mathscr{E}, t]d^2\mathscr{E}$ is the probability that the complex field amplitude has the value $\mathscr{E}$ within $d^2\mathscr{E} = \mathscr{E}d|\mathscr{E}|d\phi$ in the complex $\mathscr{E}$ plane, then the thermal (chaotic) distribution for the complex amplitude $\mathscr{E}(t)$ is a Gaussian function:

$$p^{\mathrm{th}}[\mathscr{E}] = \left(2\pi\mathscr{E}_0^2\right)^{-1/2}\exp\left[-\tfrac{1}{2}\left(|\mathscr{E}|^2/\mathscr{E}_0^2\right)\right] \quad \text{(thermal)} \qquad (72a)$$

and the coherent distributions is a delta function:

$$p^{\mathrm{coh}}[\mathscr{E}] = (|\mathscr{E}_0|/\pi)\delta^2(|\mathscr{E}|^2 - |\mathscr{E}_0|^2) \quad \text{(coherent)} \quad (72b)$$

The spatial and temporal coherence properties of radiation, measured via γ or $g^{(1)}$ or $S(\omega)$, determine the degree to which the fields at two points in space and time are able to interfere. By the use of pinholes, lenses, and filters any sample of ordinary light can be made equally monochromatic and directional as laser light. Its spatial and temporal coherence properties can be made equal to those of any laser beam. If the laser light is then suitably attenuated so that it has the same low intensity as the ordinary light, one may ask whether any important differences can remain between the ordinary light and the laser light.

Surprisingly, the answer to this fundamental question is yes. The differences can be found in higher order correlation functions, involving $\mathcal{E}^*\mathcal{E}$ to powers higher than the first, such as $\langle E^2(t)E^2(t+\tau)\rangle$. These can be referred to as intensity correlations. The first measurements of optical intensity correlation functions were made on thermal light by Brown and Twiss in the 1950s before lasers were available.

Discussions of intensity correlations are typically made with the aid of a quantity called the *degree of second-order coherence* and denoted $g^{(2)}$. Higher order degrees of coherence are also defined. If the fields are stationary only the time delays between the measurement points are significant, and one has:

$$g^{(2)}(\tau_1) = \langle I(t)I(t+\tau_1)\rangle/\langle I\rangle^2 \tag{73a}$$

$$g^{(3)}(\tau_1,\tau_2) = \langle I(t)I(t+\tau_1)I(t+\tau_2)\rangle/\langle I\rangle^3, \tag{73b}$$

and so on. Because of stationarity we can put $t = 0$ everywhere. We will deal mostly now with $g^{(2)}$ and no confusion should occur if we omit the index (2) whenever we have a single delay time τ. The assumption that the light is stationary gives $g(\tau) = g(-\tau)$. Since I is an intrinsically positive quantity, it is clear that $g(\tau)$ is positive. There is no upper limit, so g satisfies $\infty > g(\tau) \geqq 0$. In addition, for any distribution of I one has $\langle I^2\rangle \geq \langle I\rangle^2$, so obviously $g(0) \geq 1$; and one can also show that $g(0) \geq g(\tau)$.

With the aid of the thermal and coherent probability distributions given above, several higher order coherence functions can be calculated immediately. For example, if $\tau_1 = \tau_2 = \ldots = 0$, the nth order moments of the intensity are simply $\langle I^n\rangle \approx \int |\mathcal{E}|^{2n} p[\mathcal{E}] d^2\mathcal{E}$:

$$\langle I^n\rangle = n!\langle I\rangle^n = n!\bar{I}^n \quad \text{(thermal)} \tag{74a}$$

$$\langle I^n\rangle = \langle I\rangle^n = \bar{I}^n \quad \text{(coherent).} \tag{74b}$$

For large values of n these are obviously quite different, even if $\bar{I}$ is the same for two light beams, one thermal and the other coherent. The difference plays a role in multiphoton ionization experiments with $n = 2$ and larger. This is one example of a fundamental difference between ordinary light and ideal single-mode laser light.

It should be clear that in addition to thermal light and laser light there may be still other forms of light, characterized by distributions other than those in Eq. (72). Quantum theory also predicts that there can even be kinds of light beams for which the underlying probability distribution does not exist in a classical sense, for example, it may be negative over portions of the range of definition. Generalized phase space functions that play the quantum role of classical probability densities were developed by Glauber and Sudarshan in the 1960s. These are still a principal theoretical tool in studies of photon counting.

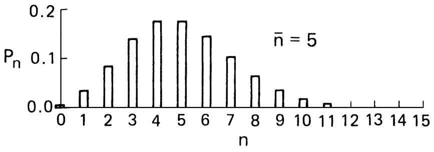


FIGURE 7 Ideal Poisson photocount distribution. [Adapted with permission from Loudan, R. (1983). "The Quantum Theory of Light," 2nd ed. Oxford Univ. Press, Oxford, UK. Copyright 1983 Oxford University Press.]

B. Photon Counting

In photon counting experiments the arrival of an individual photon is registered at a photodetector, which is essentially just a specially designed phototube that gives a signal when an arriving photon ionizes an atom at the phototube surface. A typical experiment consists of many runs of the same length T in each of which the number of photons registered by the photodetector is counted. The counts can be organized into a histogram (Fig. 7), which is interpreted as giving the probability $P_n(T)$ for counting n photons during an interval of length T.

An expression for $P_n(T)$ follows from a consideration of the photodetection process, which begins with the ionization of a single atom by a single photon at the surface of the phototube. In any event, the rate of counting is proportional to the intensity of the light beam, so one writes $\alpha I(t)dt$ for the probability of counting a photon at the time t in the interval dt. The factor α takes account of the atomic variables governing the ionization process as well as the geometry of the phototube. It was first shown by Mandel in the 1950s that $P_n(T)$ is then given by

$$P_n(T) = \left\langle e^{-\alpha\int_0^t dt' I(t')}\left[\alpha\int_0^T dt' I(t')\right]^n \Big/ n!\right\rangle, \tag{75}$$

where the average is over the variations in intensity during the (relatively long) counting intervals. Alternatively, it can be considered an average over an ensemble of identically prepared runs in which the value of $\bar{I}$ is statistically distributed in some way.

The simplest example occurs if the light intensity does not fluctuate at all, which is characteristic of an ideal single-mode laser (coherent) light beam. In this case Eq. (75) is independent of the t' average and, with $\bar{n} = \alpha\bar{I}T$.

$$P_n = e^{-n}(\bar{n})^n/n! \quad \text{(coherent),} \tag{76}$$

which is the well-known Poisson distribution. It is easily verified that $\bar{n} = \sum nP_n(T)$. That is, as its form indicates, $\bar{n}$ is the average number of photons counted in time T. It is a feature of the Poisson distribution that its dispersion is equal to its mean:

$$\langle(\Delta n)^2\rangle = \langle n\rangle^2 2 - \langle n\rangle^2 = \bar{n} \quad \text{(Poisson).} \tag{77}$$

A plot of an ideal Poisson photocount distribution is shown in Fig. 7. The Poisson distribution is also called the coherent or coherent-state distribution because it is predicted by the quantum theory of light to be applicable to a radiation field in a so-called coherent state (see Section III.C). A well-stabilized single-mode laser gives the best realization of a coherent state in practice.

It should be obvious that the same Poisson law will be found even if $\langle I(t')\rangle$ is not constant, so long as T is made great enough that all fluctuations associated with a particular interval are averaged out. The counting fluctuations associated with steady $\langle I\rangle$ are due to the discrete single-photon character of the atomic ionization event that initiates the count and are called particle fluctuations.

Although the Poisson distribution is the result for $P_n(T)$ in the simplest case, constant $\langle I(t')\rangle$, it is not the correct result for ordinary thermal light unless T is very long, in fact $2\pi\,\Delta\nu T \gg 1$ is necessary. We have posed a question at the end of the last section about differences between laser light and thermal light with equal (very narrow) bandwidths. In that case $\Delta\nu$ is very small, so we cannot automatically assume $2\pi\,\Delta\nu T \gg 1$. In fact it illustrates the point to assume the reverse. If $2\pi\,\Delta\nu T \ll 1$, we can assume $\langle I(t')\rangle$ is constant over such a short time T. However, $\langle I(t')\rangle$ can still fluctuate with t', that is, from run to run. To evaluate the average over t', which is now essentially an average over runs, we use the thermal distribution Eq. (72a) for $\mathscr{E}$, which is equivalent to the normalized exponential distribution for I

$$p(I) = (1/\bar{I})e^{-I/\bar{I}} \quad \text{(thermal)}, \tag{78}$$

since $I \approx |\mathscr{E}|^2$. Then Eq. (75) gives

$$P_n(T) = (\bar{n})^n/(1+\bar{n})^{1+n} \quad \text{(thermal)}, \tag{79}$$

which is variously called the thermal, chaotic, or Bose–Einstein distribution, and $\bar{n}$ is defined as before, $\bar{n} = \alpha\bar{I}T$, with the understanding that here $\bar{I}$ means the average over many runs, all shorter than $1/2\pi\,\Delta\nu$. The difference between the photocount distributions for thermal light under the two extreme conditions $2\pi\,\Delta\nu T \ll 1$ and $2\pi\,\Delta\nu T \gg 1$ is shown in Fig. 8. The nature of the difference between the thermal and coherent probability distributions (76) and (77), and thus of the fundamental difference between natural light and single-mode laser (coherent) light, can show up directly in the record of photocount measurements, as is clear by inspecting Figs. 7 and 8. Other photon count distributions than these two correspond to light that is somehow different from both laser light and thermal light. In quantum optics the most interesting examples are examples of purely quantum mechanical light beams.

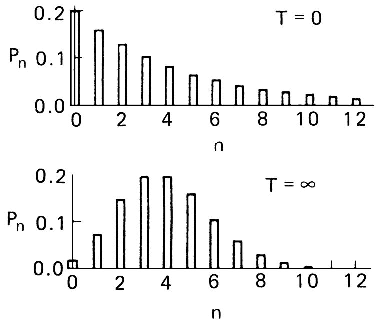


FIGURE 8 Thermal photocount records for $2p\,T\Delta n \gg 1$ and $2p\,T\Delta n \ll 1$. [Adapted with permission from Loudon, R. (1983). "The Quantum Theory of Light," 2nd ed. Oxford Univ. Press, Oxford, UK. Copyright 1983 Oxford University Press.]

C. Quantum Mechanical States of Light

The quantum theory of light assigns operators in a Hilbert space to the fields of electromagnetism. Any field $F(\mathbf{r})$, quantum or classical, can be written as a sum of plane waves (as a three-dimensional Fourier series or integral):

$$\begin{aligned} F(\mathbf{r}) &= \sum_k f_k \exp[i\mathbf{k}\cdot\mathbf{r}] \\ &= \sum_k \{ f_k \exp[i\mathbf{k}\cdot\mathbf{r}] \\ &\quad + f_{-k}\exp[-i\mathbf{k}\cdot\mathbf{r}]\}, \quad k \geqq 0 \end{aligned} \tag{80}$$

and $f_{-k} = f_k^*$ if $F(\mathbf{r})$ is a real function. If F is operator-valued real (i.e., hermitean), then the expansion coefficients f_k are operators and one writes $f_{-k} = f_k^\dagger$, where the dagger denotes hermitean adjoint in the usual way. In the case of electromagnetic radiation a given plane wave mode has the transverse electric field

$$\mathbf{E}_k(\mathbf{r}) = \hat{\mathbf{e}}N_k\big[a_k e^{i\mathbf{k}\cdot\mathbf{r}} + a_k^\dagger e^{-i\mathbf{k}\cdot\mathbf{r}}\big], \tag{81}$$

where N_k is an appropriate normalization constant. The full field is obtained by summing over k: $\mathbf{E}(\mathbf{r}) = \sum_k \mathbf{E}_k(\mathbf{r})$.

Because a_k and $a_k^\dagger$ are operators rather than numbers, $a_k^\dagger a_k \neq a_k a_k^\dagger$. This is expressed by saying that they obey the "canonical" commutation relations:

$$[a_k, a_k^\dagger] = 1, \tag{82}$$

where the square bracket means the difference $a_k a_k^\dagger - a_k^\dagger a_k$, as is usual in quantum mechanics. All other commutators among a's and $a^\dagger$'s for this mode, and all commutators of a_k or $a_k^\dagger$ with operators of all other modes vanish, for example, $[a_k, a_l] = 0$. The time dependences of these operators, and thus of the electric and magnetic fields, are determined dynamically from Maxwell's equations, which remain exactly valid in quantum theory. In the absence of charges and currents the time dependences are

$$a_k(t) = a_k e^{-i\omega_k t} \quad \text{and} \quad a_k^\dagger = a_k^\dagger e^{i\omega_k t}, \tag{83}$$

where $\omega_k = kc$.

The hermitean operator $a^\dagger a$ (we drop the mode index k temporarily) has integer eigenvalues and is called the photon number operator:

$$a^\dagger a|n\rangle = n|n\rangle, \quad n = 0, 1, 2, \ldots, \tag{84}$$

and the eigenstate $|n\rangle$ is called a Fock state or *photon number state*. The operators $a^\dagger$ and a are called the photon *creation operator* and *destruction operator* because of their effect on photon number states:

$$a^\dagger|n\rangle = \sqrt{(n+1)}|n+1\rangle,$$

and (85)

$$a|n\rangle = \sqrt{n}|n-1\rangle.$$

That is, $a^\dagger$ and a respectively increase and decrease by 1 the number of photons in the field state. The photon number states are mutually orthonormal: $\langle m|n\rangle = \delta_{mn}$, and they form a complete set for the mode, and all other states can be expressed in terms of them. They have very attractive simple properties. For example, the photon number is exactly determined, that is, the dispersion of the photon number operator $a^\dagger a$ is zero in the state $|n\rangle$ for any n:

$$\begin{aligned}\langle(\Delta n)^2\rangle &= \langle n|(a^\dagger a)^2|n\rangle - \langle n|\, a^\dagger a|\rangle^2\\ &= n^2 - n^2 = 0.\end{aligned}$$

However, the nice properties of the $|n\rangle$ states are in some respects not well suited to the most common situations in quantum optics. For example, loosely speaking, their phase is completely undefined because their photon number is exactly determined. As a consequence, the expected value of the mode field is exactly zero in a photon number state. That is $\langle n|\mathbf{E}(\mathbf{r})|n\rangle = 0$, because Eq. (85) and the orthogonality property together give $\langle n|a|n\rangle = \langle n|a^\dagger|n\rangle = 0$.

In most laser fields there are so many photons per mode that it is difficult to imagine an experiment to count the exact number of them. Thus, a different kind of state, called a *coherent state*, is usually more appropriate to describe laser fields than the photon number states. These states $|\alpha\rangle$ are right eigenstates of the photon destruction operator a:

$$a|\alpha\rangle = \alpha|\alpha\rangle. \tag{86}$$

They are normalized so that $\langle\alpha|\alpha\rangle = 1$. Since a is not hermitean, its eigenvalue α is generally complex. The expected number of photons in the mode in a coherent state is $\langle n\rangle = \langle a^\dagger a\rangle = \langle\alpha|a^\dagger a|\alpha\rangle = |\alpha|^2$. Thus, α can be interpreted loosely as $\sqrt{\langle n\rangle}$, the square root of the mean number of photons in the mode. The number of photons in the mode is not exactly determined in the state $|\alpha\rangle$. This can be seen by computing the dispersion in the number:

$$\begin{aligned}\langle(\Delta n)^2\rangle &= \langle\alpha|(a^\dagger a)^2|\alpha\rangle - \langle\alpha|a^\dagger a|\alpha\rangle^2\\ &= \langle\alpha|(a^\dagger(a^\dagger a-1)a|\alpha\rangle - \langle\alpha|a^\dagger a|\rangle^2\\ &= \langle\alpha|(a^\dagger a|\alpha\rangle = |\alpha|^2 = \langle n\rangle.\end{aligned} \tag{87}$$

Thus, the dispersion is equal to the mean number, a property already noticed for the Poisson distribution (77). If $\langle n\rangle \gg 1$, then the relative dispersion $\langle(\Delta n)^2\rangle/\langle n\rangle^2$ is very small, and the number of photons in the state is well determined in a relative sense. At the same time, a relatively well-defined phase can also be associated with the state.

The clearest interpretation of the amplitude and phase associated with a coherent state $|\alpha\rangle$ is obtained by computing the expected value of $\mathbf{E}$ in the state $|\alpha\rangle$. Since $\langle\alpha|a^\dagger|\alpha\rangle = \langle\alpha|a|\alpha\rangle^* = \alpha^*$, one easily determines that

$$N_k = \sqrt{(\hbar\omega_k/2\varepsilon_0 V)}, \tag{88}$$

where V is the mode volume, and then N_k can be interpreted loosely as the electric field amplitude associated with one photon. This shows that $N_k\alpha = \sqrt{(\hbar\omega_k/2\varepsilon_0 V)}\alpha$ is the mean amplitude of the expected electric field, or what is sometimes called the classical field amplitude. The phase of α thus determines the phase of the field described by the state $|\alpha\rangle$.

The coherent states $|\alpha\rangle$ can be expressed in terms of the photon number states:

$$|\alpha\rangle = \exp[-|\alpha|^2/2]\sum_m \left[\frac{a^m}{\sqrt{m!}}\right]|m\rangle, \tag{89}$$

where the sum runs over all integers $m \geqq 0$. From Eq. (89) one can compute the probability $p_n(\alpha)$ that a given coherent state $|\alpha\rangle$ contains exactly n photons. According to the principles of quantum theory this is given by $|\langle n|\alpha\rangle|^2$, and we find:

$$p_n(\alpha) = e^{-\langle n\rangle}\langle n\rangle^n/n!, \tag{90}$$

which is exactly the "purely random" Poisson probability distribution shown in Fig. 7. This is interpreted as meaning that photon-counting measurements of a radiation field in the coherent state $|\alpha\rangle$ would find exactly n photons with probability (90).

The photon creation and destruction operators are not hermitean and are generally considered not observable, but their real and imaginary parts (essentially the electric and magnetic field strengths, or in other terms, the in-phase and quadrature components of the optical signal) are in principle observable. Thus, one can introduce the definitions

$$a = \tfrac{1}{2}(a_1 - ia_2), \quad a^\dagger = \tfrac{1}{2}(a_1 + ia_2) \tag{91}$$

and their inverses

$$a_1 = a + a^\dagger, \quad a_2 = i(a - a^\dagger). \tag{92}$$

What are the quantum limitations on measurement of the hermitean operators a_1 and a_2? They must, of course, obey the Heisenberg *uncertainty relation* $\langle(\Delta a_1)^2\rangle\langle(\Delta a_2)^2\rangle \geqq |\frac{1}{2}\langle[a_1, a_2]\rangle|^2$, where $\langle\ldots\rangle$ indicates expectation value in any given quantum state, and

$\Delta a_1 \equiv a_1 - \langle a_1 \rangle$. Since $[a, a^\dagger] = 1$, it follows from Eq. (91) that $[a_1, a_2] = -2i$, so the Heisenberg relation for a_1 and a_2 is:

$$\langle (\Delta a_1)^2 \rangle \langle (\Delta a_2)^2 \rangle \geq 1. \tag{93}$$

A coherent state $|\alpha\rangle$ can be shown to produce the minimum simultaneous uncertainty in a_1 and a_2. That is, $\langle \alpha | (\Delta a_1)^2 | \alpha \rangle = 1$ and $\langle \alpha | (\Delta a_2)^2 | \alpha \rangle = 1$. Both a_1 and a_2 are therefore said to reach the *quantum limit* of uncertainty in a coherent state.

However, it is only the product of the operator dispersions that is constrained by the Heisenberg uncertainty relation, and there is no fundamental reason why either a_1 and a_2 could not have a dispersion equal to, $\mu \ll 1$, so long as the other had a dispersion at least as large as $1/\mu \gg 1$. A quantum state of the radiation field that permits one of the components of the destruction operator to have a dispersion smaller than the quantum limit is said to be a *squeezed state*. Squeezed states could in principle provide the ability to make ultraprecise measurements such as are projected for gravity wave detection. A squeezed state of radiation was first generated and measured by Slusher and others in 1985.

IV. QUANTUM INTERACTIONS AND CORRELATIONS

It should be remembered that the highly successful semiclassical version of the quantum theory of light (Sections I and II) does not ignore quantum principles, or put $\hbar \to 0$, but it does ignore quantum fluctuations and correlations. It is a theory of coupled quantum expectation values.

In the following sections of a number phenomena depending directly on quantum fluctuations and correlations are described. None of them have been successfully treated by the semiclassical theory or by any other theory than the generally accepted and fully quantized version of the quantum theory of light. For this reason the observation of these and similar quantum optical effects can offer a means of testing the accepted theory.

Such tests are of great interest for two related reasons. Because the quantum theory of light (or quantum electrodynamics) is the most carefully studied quantum theory, and because it serves as a fundamental guide to all field theories, it plays a key role in our present understanding of quantum principles and should be tested as rigorously as possible. Moreover, tests of the effects described below play a special role because they bear on the theory in a different way compared with traditional tests, such as high-precision measurements of the Lamb shift, the fine structure constant, the Rydberg, and the electron's anomalous moment.

A. Fully Quantized Interactions

The electric field is mainly responsible for optical interactions of light with matter, and the magnetic field plays a subsidiary role, becoming significant only in situations involving magnetic moments or relativistic velocities. The most important light–matter interactions is the direct coupling of electric dipoles to the radiation field through the interaction energy $-\mathbf{d} \cdot \mathbf{E}$.

A systematic description of the fully quantized interactions of quantum optics begins with the total energy of the atom H_A and the radiation H_R and their energy of interaction:

$$H = H_\mathrm{A} + H_\mathrm{R} - \mathbf{d} \cdot \mathbf{E}. \tag{94}$$

In the quantum theory the atomic, radiation, and interaction energies are given by the Hamiltonian operators

$$H_\mathrm{A} = \sum_j E_j |j\rangle\langle j| \tag{95a}$$

$$H_\mathrm{R} = \sum_k \hbar\omega_k a_k^\dagger a_k \tag{95b}$$

$$-\mathbf{d} \cdot \mathbf{E} = -\sum_i \sum_j \sum_k \hbar f_{ij}^k \{a_k^\dagger + a_k\} |i\rangle\langle j|. \tag{95c}$$

Here E_j and $|j\rangle$ are the quantized energies and eigenstates of the atom including level-shifting and level-splitting due to static fields that give rise to Zeeman and Stark effects, etc. The dipole coupling constant is $\hbar f_{ij}^k = N_k \hat{\mathbf{e}} \cdot d_{ij}$, in the notation of Eqs. (9) and (39). Also, $a_k^\dagger$ and a_k are the photon creation and destruction operators introduced in Section III.C.

The two-level version of this Hamiltonian is the most used in quantum optics. It is obtained by restricting the sums over i and j to the values 1 and 2. It is not necessary that $|1\rangle$ and $|2\rangle$ be the two lowest energy levels. In photoionization, which is the physical process underlying the operation of photon counters, the upper state is not even a discreate state but lies in the continuum of energies above the ionization threshold (see Section IV.B). In its two-level version H becomes

$$\begin{aligned} H = {} & E_1|1\rangle\langle 1| + E_2|2\rangle\langle 2| + \sum_k \hbar\omega_k a_k^\dagger a_k \\ & - \sum_k \hbar f_{12}^k \{a_k^\dagger + a_k\}\{\sigma + \sigma^\dagger\}, \end{aligned} \tag{96}$$

where $\sigma = |1\rangle\langle 2|$ and $\sigma^\dagger = |2\rangle\langle 1|$, and f_{12}^k has been taken to be real for simplicity. Note that σ has the effect of a lowering transition when it acts on the two-level atomic state $|\Psi\rangle = C_1|1\rangle + C_2|2\rangle$. That is, $\sigma|\Psi\rangle = C_2|1\rangle$, so σ takes the amplitude C_2 of the upper state $|2\rangle$ and assigns it to the lower state $|1\rangle$. By the same argument $\sigma^\dagger$ causes a raising transition.

The term $a_k^\dagger \sigma^\dagger$ in Eq. (96) is difficult to interpret because it has $\sigma^\dagger$ raising the atom into its upper state together with $a_k^\dagger$ creating a photon. One expects photon creation to be associated only with a lowering of the atomic state. The term $a_k \sigma$ presents similar difficulties. It can be shown, however, that these two terms are the source of the very rapid oscillations $\exp[\pm 2i\omega t]$ which were discussed above Eq. (12) and were eliminated by the rotating wave approximation (RWA). The adoption of the RWA eliminates them here also. With the RWA and the convenient convention that $E_1 = 0$ (and therefore that $E_2 = \hbar\omega_{21}$), the working two-level Hamiltonian is

$$H = \frac{1}{2}\hbar\omega_{21}(\sigma_z + 1) + \sum_k \hbar\omega_k a_k a_k^\dagger - \sum_k \hbar\, f_{12}^k \{a_k^\dagger \sigma + \sigma^\dagger a_k\}. \tag{97}$$

The operators $\sigma, \sigma^\dagger$, and σ_z are closely related to the 2×2 Pauli spin matrices:

$$|1\rangle\langle 2| = \sigma \to \tfrac{1}{2}(\sigma_x - i\sigma_y) = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \tag{98a}$$

$$|2\rangle\langle 1| = \sigma^\dagger \to \tfrac{1}{2}(\sigma_x - i\sigma_y) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \tag{98b}$$

$$|2\rangle\langle 2| - |1\rangle\langle 1| \to \sigma_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}. \tag{98c}$$

One can easily confirm that the matrix representation of σ is a "lowering" operator in the two-dimensional space with basis vectors

$$|2\rangle \to \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{and} \quad |1\rangle \to \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{98d}$$

and $\sigma^\dagger$ is represented by the corresponding "raising" operator. The σ's obey the commutator relations:

$$[\sigma, \sigma_z] = 2\sigma, \quad [\sigma^\dagger, \sigma_z] = -2\sigma^\dagger, \quad [\sigma^\dagger, \sigma] = \sigma_z. \tag{99}$$

Changes in the radiation field occur as a result of emission and absorption of photons during transitions in the atom. One consequence is that a_k obeys an equation obtained from the Heisenberg equation $i\hbar\partial O/\partial t = [O, H]$ that is valid for all operators O:

$$\partial a_k/\partial t = -i\omega_k a_k + i f_k \sigma. \tag{100}$$

Here we have simplified the notation, rewriting f_{12}^k as f_k. The solution, of Eq. (100) is:

$$a_k(t) = a_k(0)e^{-i\omega_k t} + i f_k \int_0^t dt' e^{-i\omega_k(t-t')}\sigma(t'). \tag{101}$$

The first term represents photons that are present in the mode k but not associated with the two-level atoms (e.g., from a distant laser), and the second term represents the photons associated with a transition in the two-level atom.

The σ operators also change in time in the course of emission and absorption processes. Their time dependence is also determined by Heisenberg's equation and one finds the equations:

$$d\sigma/dt = -i\omega_{21}\sigma - i\sum_k f_k a_k \sigma_z \tag{102a}$$

$$d\sigma^\dagger/dt = i\omega_{21}\sigma^\dagger + i\sum_k f_k a_k^\dagger \sigma_z \tag{102b}$$

$$d\sigma_z/dt = 2i\sum_k f_k(\sigma^\dagger a_k - a_k^\dagger \sigma). \tag{102c}$$

These are the operator equations underlying the semiclassical equations for the Bloch vector components given in Eq. (19).

The correspondence between the quantum and semiclassical sets of variables is

$$\sigma \to \tfrac{1}{2}(u - iv)e^{-i\omega t} \tag{103a}$$

$$\sigma^\dagger \to \tfrac{1}{2}(u + iv)e^{i\omega t} \tag{103b}$$

$$\sigma_z \to w. \tag{103c}$$

This identification is precise if the radiation field is both intense and classical enough. This means that one retains only one mode and replaces a_k and $a_k^\dagger$ by their coherent state expectation values $\alpha = \alpha_0\, e^{-i\phi}$ and $\alpha_0^*\, e^{i\phi}$. Then Eqs. (102) are identical with Eq. (19) under the previous assumptions, that is, the field is quasi-monochromatic so $\phi = \omega t$, and $N_k \alpha$ is the slowly varying electric field expectation value, which is interpreted as the classical field amplitude. Thus, one has

$$2\, f_k \alpha_0 \to 2d\mathscr{E}_0/\hbar = \chi. \tag{104}$$

The part of a_k that depends on σ in Eq. (100) acts as a radiation reaction field when substituted into Eq. (102). It causes damping and a small frequency shift in the atomic operator equations even if there are no external photons present and thus is associated with spontaneous emission. The damping constant is exactly the correct Einstein A coefficient for the transition, because the A coefficient is a two-level parameter.

The frequency shift is only a primitive two-level version of a more general many level radiative correction such as the Lamb shift. One observes here a natural limitation of any two-level model. It is intrinsically incapable of dealing with any precision with effects, such as radiative level shifts, that depend strongly on the contributions of many levels. However, in the cases of interest described in the remainder of Section IV, the effects of other levels are negligible and the numerical value of the frequency shift is irrelevant. It can be assumed to be included in the definition of ω_{21}.

B. Quantum Light Detection and Statistics

The quantum theory of light detection is based on the quantum theory of photoionization because photon counters are triggered by an ionizing absorption of a photon. Photoionization is a weak field phenomenon because the effective γ is so large [recall Eq. (45)]. Thus, perturbative methods are adequate and one computes the absolute square of the ionization matrix element, in this case given by $\langle F| - e\mathbf{r}\cdot\mathbf{E}(\mathbf{r})|I\rangle$, where $|I\rangle$ and $|F\rangle$ are the initial and final states of the photoionization. The initial state consists of an atom in its ground state, described by the electronic orbital function $\phi_0(\mathbf{r})$, and the initial state of the radiation field $|\Psi\rangle$. The final state consists of the atom in an ionized state described by an electronic orbital function $\phi_f(\mathbf{r})$ appropriate to a free electron with energy above the ionization threshold, and another state of the radiation field $|\Psi'\rangle$. Then the matrix element becomes

$$\begin{aligned}\langle F| - e\mathbf{r}\cdot\mathbf{E}|I\rangle &\\ &= -\int \phi_f^*(\mathbf{r})e\mathbf{r}\cdot\langle\Psi'|\mathbf{E}(\mathbf{r})|\Psi\rangle\phi_0(\mathbf{r})d^3r\\ &= -\mathbf{d}_{f0}\cdot\sum_k \hat{\mathbf{e}}_k N_k\langle\Psi'|a_k|\Psi\rangle, \qquad (105)\end{aligned}$$

where $\mathbf{d}_{f0}$ is the so-called dipole matrix element for the $0 \rightarrow f$ transition in the atom. We have also made the "dipole" approximation, in which $\exp[i\mathbf{k}\cdot\mathbf{r}] \approx \exp[i0] \approx 1$ over the entire effective range of the matrix element integral [recall the discussion following Eq. (9)].

Only the part of $\mathbf{E}$ that lowers the photon number of the field, namely the "a" part, is effective in ionization, essentially because ionization is a photon-absorption process. In addition we can take a single k value if the incident light is monochromatic. Thus, the ionization rate depends on $f_k^2|\langle\Psi'|a_k|\Psi\rangle|^2$, where $f_k = |\mathbf{d}_{f0}\cdot\hat{\mathbf{e}}_k|N_k$. The actual final states of the atom and of the field are never completely observed, and all the unobserved features must be allowed for, that is, included by summation:

$$\begin{aligned}\text{rate} &\approx f_k^2\sum_{\Psi'}\langle\Psi|a_k^\dagger|\Psi'\rangle\langle\Psi'|a_k|\Psi\rangle\\ &\approx f_k^2\langle\Psi|a_k^\dagger a_k|\Psi\rangle \approx \langle\Psi|a_k^\dagger a_k ¢\Psi\rangle,\end{aligned}$$

since $\sum_{\Psi'}|\Psi'\rangle\langle\Psi'| = 1$ for a complete set of final states. In this expression for the ionization rate we write "$\approx$" instead of "$=$" because we are really interested here only in the effects of field quantization on the rate, not the exact numerical value of the rate. If the radiation field is quantum mechanical we do not know perfectly the properties of the incident light, and these properties must be averaged. This average over the properties of the incident light is the same average discussed from a classical standpoint in Sections III.A and III.B. Thus, we finally have

$$\text{rate} \approx \langle a_k^\dagger a_k\rangle, \qquad (106)$$

where the angular brackets now mean an average over the initial field, that is, the quantum mechanical expectation value in state $|\Psi\rangle$.

The significance of Eq. (106) is in the ordering of the field operators. The nature of the photoionization process mandates that they be in the given order and not the reverse. Since $a^\dagger a$ is not equal to $aa^\dagger$ for a quantum field, the order makes a difference. The order given, in which the destruction operator is to the right, is called *normal order*. Photoionization is a normally ordered process by its nature, and therefore so is photodetection and photon counting. This has fundamental consequences for quantum statistical measurements, as we now explain.

Let us consider the degree of second-order coherence $g^{(2)}$ in quantum theory. This was written in Eq. (73a) as an intensity correlation: $g^{(2)}(\tau) = \langle I(t)I(t+\tau)\rangle/\langle I\rangle^2$. Because of the normally ordered character of photoionization, if $g^{(2)}$ is measured with photodetectors as usual, its correct definition according to the quantum theory of ionization is normally ordered:

$$g^{(2)}(\tau) = \frac{\left\langle E^{(-)}(t)E^{(-)}(t+\tau)E^{(+)}(t+\tau)E^{(+)}(t)\right\rangle}{\left\langle E^{(-)}E^{(+)}\right\rangle\left\langle E^{(-)}E^{(+)}\right\rangle}. \qquad (107)$$

If photon fields were really classical, and these quantum mechanical fine points were unnecessary, then the ordering would make no difference, since the fields $E^{(\pm)}$ would be numbers, not operators, and the original expression for $g^{(2)}$ would be recovered. However, we now exhibit the effects of these quantum differences in a few specific cases.

In the case of a single-mode field, there are no time dependences and $g^{(2)}(\tau) = g^{(2)}(0)$ simplifies to

$$g^{(2)}(0) = \langle a^\dagger a^\dagger aa\rangle/\langle a^\dagger a\rangle^2, \qquad (108)$$

which we evaluate in Table II. Among these examples the Fock state is special because its $g^{(2)}$ violates the condition $g^{(2)}(0) \geq 1$, which is one of the classical inequalities given below Eqs. (73). The Fock state is therefore an example of a state of the radiation field for which the quantum and classical theories make strikingly different predictions. It has not yet been possible to study a pure Fock state of more than one photon in the laboratory.

Photon bunching is a term that refers to the fact that photon beams exist in which photons are counted with statistical fluctuations greater than would be expected on the basis of purely random (that is, Poisson) statistics. In

TABLE II Second-Order Degree of Coherence for Single-Mode Quantum Mechanical Fields in Different States

Quantum field state	Value of $g^{(2)}(0)$
Vacuum state $\|0\rangle$	0
Fock state $\|n\rangle$	0, if $n < 2$
	$1 - 1/n, \quad \text{if } n \geq 2$
Coherent state $\|\alpha\rangle$	1
Thermal state	2

fact, almost any ordinary beam (thermal light) will have this property, and this is reflected in that $g^{(2)} > 1$ for thermal light. Photon bunching therefore arises from the Bose–Einstein distribution [Eq. (39)]. A coherent state with its Poisson statistics is purely random and does not exhibit bunching.

A qualitative classical explanation of photon bunching is sometimes made by saying that light from any natural source arises from broadband multimode photon emission by many independent atoms. There are naturally random periods of constructive and destructive interference among the modes, giving rise to large intensity "spikes," or "bunches" of photons, in the light beam. Unbunched light comes from a coherently regulated collection of atoms, such as from a well-stabilized single-mode laser. From this point of view, unbunched coherent light is optimally ordered.

However, *photon antibunching* can also occur. There are "antibunched" light beams in which photons arrive with lower statistical fluctuations than predicted from a purely coherent beam with Poisson statistics. Antibunched light beams have values of $g^{(2)} < 1$, in common with a pure Fock state beam, and are therefore automatically nonclassical light beams.

The first observation of an antibunched beam with $g^{(2)} < 1$ was accomplished by Mandel and others in 1977 in an experiment with two-level atoms undergoing resonance fluorescence. Antibunching occurs in such light for a very simple reason. A two-level atom "regulates" the occurrence of pairs of emitted photons very severely, even more so than the photons are regulated in a single-mode laser. A second fluorescent photon cannot be emitted by the same two-level atom until it has been re-excited to its upper level by the absorption of a photon from the main radiation mode. Thus, a high Rabi frequency χ permits the degree of second-order coherence $g^{(2)}(\tau) = \langle a^\dagger a^\dagger(\tau) a(\tau) a\rangle / \langle a^\dagger a\rangle^2$ to be nonzero after a relatively short value of the time delay τ, but $g^{(2)}$ is strictly zero for $\tau = 0$. A graph showing the experimental observation is given in Fig. 9.

The significance of photon statistics and photon counting techniques in quantum optics and in physics is clear.

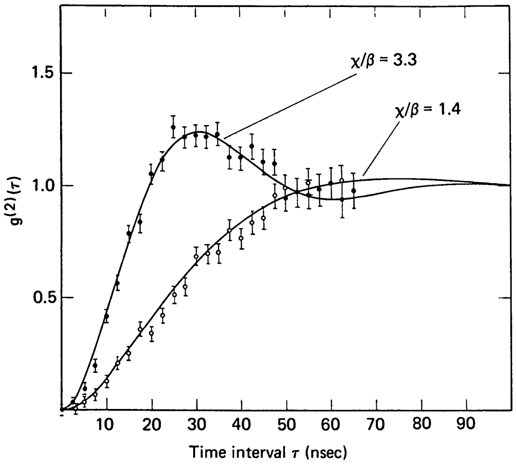


FIGURE 9 Curves of the second-order degree of coherence, showing photon antibunching. [Reprinted with permission from Dagenais, M., and Mandel, L. (1978). Investigation of two-time correlations in photon emission from a single atom. *Phys. Rev. A* **18,** 2217–2228.]

They permits a direct examination of some of the fundamental distinctions between the quantum mechanical and classical concepts of radiation.

C. Superradiance

Nonclassical photon counting statistics arise from the multiphoton correlations inherent in specific states of the light field. Similarly, multiatom quantum correlations can give rise to unusual behavior by systems of radiating atoms, as was pointed out by Dicke in 1954. The most dramatic behavior of this kind is called *Dicke superradiance* or *superfluorescence*.

Multiatom correlations can exist in N-atom systems even if N is as small as $N = 2$. A pair of two-level atoms labeled a and b can have quantum states made of linear combinations of the elementary two-atom states

$$|a+\rangle * |b+\rangle \tag{109a}$$

$$|a+\rangle * |b-\rangle \tag{109b}$$

$$|a-\rangle * |b+\rangle \tag{109c}$$

$$|a-\rangle * |b-\rangle. \tag{109d}$$

Here the sign $*$ indicates a tensor product of the vector spaces of the two atoms, and $+$ and $-$ designate the upper and lower energy levels, with energies E_2 and E_1, in each of the identical atoms. The two middle states Eq. (109b and c) are degenerate, with total energy $E_1 + E_2 = E_2 + E_1$. It is useful to define the "singlet" and "triplet" states $\|S\rangle$ and $\|T\rangle$ that are linear combinations of the degenerate states as follows:

$$\|S\rangle = (1/\sqrt{2})\{|\mathrm{a}\rangle * |\mathrm{b}-\rangle - |\mathrm{a}-\rangle * |b+\rangle\} \qquad (110\mathrm{a})$$

$$\|T\rangle = (1/\sqrt{2})\{|\mathrm{a}\rangle * |\mathrm{b}-\rangle + |\mathrm{a}-\rangle * |b+\rangle\} \qquad (110\mathrm{b})$$

in analogy to singlet and triplet combinations of two spin $-\frac{1}{2}$ states.

The two-atom interaction with the radiation field is through $\mathbf{d}\cdot\mathbf{E}$ as in the one-atom case [recall Eq. (95)], and here both atoms contribute a dipole moment:

$$\begin{aligned}\hat{\mathbf{e}}\cdot\mathbf{d} &= \hat{\mathbf{e}}\cdot[\mathbf{d}(\mathrm{a})+\mathbf{d}(\mathrm{b})]\\ &= \hat{\mathbf{e}}\cdot(\mathbf{d}_{21})_{\mathrm{a}}\{|\mathrm{a}+\rangle\langle\mathrm{a}-| + |\mathrm{a}-\rangle\langle\mathrm{a}+|\}\\ &\quad + \hat{\mathbf{e}}\cdot(\mathbf{d}_{21})_{\mathrm{b}}\{|\mathrm{b}+\rangle\langle\mathrm{b}-| + |\mathrm{b}-\rangle\langle\mathrm{b}+|\}\\ &= d\{|\mathrm{a}+\rangle\langle\mathrm{a}-| + |\mathrm{a}-\rangle\langle\mathrm{a}+|\\ &\quad + |\mathrm{b}+\rangle\langle\mathrm{b}-| + |\mathrm{b}-\rangle\langle\mathrm{b}+|\}, \qquad (111)\end{aligned}$$

where we have taken equal matrix elements: $\hat{\mathbf{e}}\cdot(\mathbf{d}_{21})_{\mathrm{a}} = \hat{\mathbf{e}}\cdot(\mathbf{d}_{21})_{\mathrm{b}} = d$.

The main features of superradiance lie in the fact that the two-atom dipole interaction causes transitions between the various states of the system at different rates, and the reason for the difference is the existence of greater internal two-atom coherence in the case of the triplet state. Suppose that the two-atom system is fully excited into the state $\|+2\rangle = |a+\rangle * |b+\rangle$, which has energy $2E_2$, and then emits one photon. The system must drop into a state with energy $E_2 + E_1$. From this state it can decay further by emission of a second photon to the ground state $\|-2\rangle = |a-\rangle * |b-\rangle$ which has energy $2E_1$.

According to the Fermi Golden Rule, the rate of these transitions depends on the square of the interaction matrix element between initial and final states, $\langle F|\mathbf{d}\cdot\mathbf{E}|\mathrm{I}\rangle$, summed over all possible final states. We consider the second transition, so there is only one possible final atomic state, namely $\|-2\rangle$. The matrix elements factor into an atomic part and a radiation part. We can use Eq. (81) in the dipole approximation to write

$$\langle F|\mathbf{d}\cdot\mathbf{E}|\mathrm{I}\rangle = \sum_k \langle F_A|\mathbf{d}\cdot\hat{\mathbf{e}}_k|I_A\rangle\langle F_R\rangle|N_k(a_k + a_k^\dagger)|I_R\rangle.$$

Since superradiance deals only with spontaneous emission, the initial radiation state is the empty or vacuum state. Thus, the field contribution to the matrix element comes just from $a_k^\dagger$ and is the same in all cases and not interesting.

The various possible atomic matrix elements are

$$\langle -2\|\mathbf{d}\cdot\hat{\mathbf{e}}_k|\mathrm{a}+\rangle * |\mathrm{b}-\rangle = d \qquad (112\mathrm{a})$$

$$\langle -2\|\mathbf{d}\cdot\hat{\mathbf{e}}_k|\mathrm{a}-\rangle * |\mathrm{b}+\rangle = d \qquad (112\mathrm{b})$$

$$\langle -2\|\mathbf{d}\cdot\hat{\mathbf{e}}_k\|0, S\rangle = 0 \qquad (112\mathrm{c})$$

$$\langle -2\|\mathbf{d}\cdot\hat{\mathbf{e}}_k\|0, T\rangle = (\sqrt{2})d, \qquad (112\mathrm{d})$$

so their squares have the relative values $1:1:0:2$. That is, the triplet initial state can radiate twice as strongly as either of the original degenerate states, and the singlet state cannot radiate at all. Both the triplet and singlet states are said to be two-body "cooperative" states because they cannot be factored into one-atom states.

It is tempting to interpret these results by saying that the triplet state radiates more rapidly because it has a larger dipole moment than the others and that the singlet state has no dipole moment. Such an interpretation is in the spirit of semiclassical radiation theory, as described in Section I.C, where the expectation values of quantum operators are treated as if they are classical variables. This interpretation has a number of useful features, but must also contain serious flaws, because the observation that it is based on is not true. A calculation of the dipole expectation value shows $\langle\Psi\|\mathbf{d}\|\Psi\rangle = 0$, where $\|\Psi\rangle$ can be any of the two-atom states above, including the rapidly radiating triplet state.

A state with a large dipole moment expectation does exist, namely the factored state

$$\|\Psi_d\rangle = \tfrac{1}{2}\{|\mathrm{a}+\rangle + |\mathrm{a}-\rangle\} * \{|\mathrm{b}+\rangle + |b-\rangle\}. \qquad (113\mathrm{a})$$

This state is actually an eigenstate of the total dipole operator:

$$\begin{aligned}\hat{\mathbf{e}}\cdot d\|\Psi_d\rangle &= [\hat{\mathbf{e}}\cdot(\mathbf{d}_{21})_{\mathrm{a}}\{|\mathrm{a}+\rangle\langle\mathrm{a}-| + |\mathrm{a}-\rangle\langle\mathrm{a}+|\}\\ &\quad + \hat{\mathbf{e}}\cdot(\mathbf{d}_{21})_{\mathrm{b}}\{|\mathrm{b}+\rangle\langle\mathrm{b}-| + |\mathrm{b}-\rangle\langle\mathrm{b}+|\}]\|\Psi\rangle\\ &= 2d\|\Psi\rangle, \qquad (113\mathrm{b})\end{aligned}$$

so one has $\langle\Psi_d\|\mathbf{d}\cdot\hat{\mathbf{e}}\|\Psi_d\rangle = 2d$. This state is also predicted to radiate strongly.

The extrapolation of these predictions to an N-atom system leads to the prediction of a very large N-atom emission intensity I_N. Related predictions are that the N-atom cooperative process begins with a relatively slow buildup. After a delay of average length $\delta\tau_N$, an ultrashort burst of radiation of duration τ_N occurs. In the ideal case one predicts

$$I_N \approx N^2\hbar\omega_{21}\big/\tau_1 \qquad (114\mathrm{a})$$

$$\tau_N \approx \tau_1/N \qquad (114\mathrm{b})$$

$$\delta\tau_N \approx \tau_1\ln(N), \qquad (114\mathrm{c})$$

where τ_1 is the single-atom radiative lifetime: $1/\tau_1 \equiv A$.

If one imagines even small collections of atoms with $N \approx 10^{12}$, then N^2 is impressively very much bigger than N and the term superradiance is indeed apt. If $A \approx 10^8\ \mathrm{s}^{-1}$ is taken as typical for 2 eV optical transitions, then $N \approx 10^{12}$ suggests that 2×10^{12} eV of energy can be expected at the rate $10^{20}\ \mathrm{s}^{-1}$ for a purely spontaneous power output of $2\times10^{32}\ \mathrm{eV\ s}^{-1} \approx 3\times10^{11}$ W.

Other aspects of superradiance are equally interesting on fundamental grounds. For example, which of the two large dipole states, $\|T\rangle$ or $\|\Psi_d\rangle$, is actually responsible for superradiance? They both predict $I_N \approx N^2\hbar\omega_{21}$, but their correlation properties are completely different, in somewhat the same way that a photon number state $|n\rangle$ and a coherent state $|\alpha\rangle$ have very different correlation properties even if they predict the same mode energy $|\alpha|^2\hbar\omega = n\hbar\omega$. Consider only the fluctuations in $\mathbf{d}$ itself for the two states. If one calculates the expectation of the dispersion of $\Delta\mathbf{d}^2 \equiv \langle[\mathbf{d}\cdot\hat{\mathbf{e}} - \langle\mathbf{d}\cdot\hat{\mathbf{e}}\rangle]^2\rangle$, one finds:

$$\langle\Psi_d\|\Delta\mathbf{d}^2\|\Psi_d\rangle = 0 \tag{115}$$

$$\langle T\|\Delta\mathbf{d}^2\|T\rangle = d^2. \tag{116}$$

One can infer that radiation from the state $\|T\rangle$ can be expected to exhibit strong fluctuations of a kind completely absent in radiation from the state $\|\Psi_d\rangle$.

The fluctuations predicted from the state $\|T\rangle$ are consistent with the fact that it is exactly the state connected directly to the initial fully excited state $\|{+2}\rangle$ by the total dipole operator $\mathbf{d}$. That is, $\hat{\mathbf{e}}\cdot d\| + 2\rangle = (\sqrt{2})d\|T\rangle$. The fluctuations can be associated with the quantum uncertainty in the emission time of the first photon. Such fluctuations will influence all subsequent evolution, and if $N \gg 1$, they can be regarded as an example of a *macroscopic quantum fluctuation*, that is, a fluctuation with quantum mechanical origins that achieves direct macroscopic observability.

For a period of years, superradiance was a controversial and unobserved phenomenon. The intense and highly directional light beam predicted for the effect suggests that each emitted photon contributes to a spontaneous radiation field, which helps to stimulate the emission of further photons. Such a self-reactive process would provide a feedback analogous to that provided by mirror reflections in a laser cavity. It has been suggested that the physical origins of superradiance and laser emission are in fact the same thing. Important differences exist, however. During laser action, the dipole coherence of an individual atom is interrupted by collisions extremely frequently. The incoherent adiabatic solution for ρ_{21} is a quite satisfactory element of all laser theories (recall Section II.G). By contrast, in ideal Dicke superradiance all N-atom dipole coherence is fully preserved during the entire radiation process.

The experimental observation of superradiance was first achieved in the 1970s by Feld, Haroche, Gibbs, and others. Agreement has been found with the correlated state predictions, particularly with the statistical nature of the delay time fluctuations, and there is no longer any controversy over its existence. However, important questions about quantum and propagation effects on the spatial coherence properties of superradiance remain open.

D. Two-Level Single-Mode Interaction

We have emphasized that beginning with Einstein's reconsideration of Planck's radiation law, the most fundamental interacting system in quantum optics is a single two-level atom coupled to a single mode of the radiation field. This interaction was described semiclassically in Section II.A, and the quantum mechanical origin of the semiclassical equations was explained in Section IV.A. Fully quantum mechanical studies of the quantum coherence properties of this simplest interacting system were initiated by Jaynes in the 1950s. Some of the differences between general quantum and semiclassical theories have been clarified in this context.

When only one mode is significant the Hamiltonian [Eq. (97)] can be reduced to:

$$\begin{aligned} H_{\mathrm{JC}} = {} & \tfrac{1}{2}\hbar\omega_{21}(\sigma_z + 1) + \hbar\omega a^\dagger a \\ & + \hbar\lambda(a^\dagger\sigma + \sigma^\dagger a). \end{aligned} \tag{117}$$

Here $\lambda = f_{12}^k$, which is assumed real for simplicity. Remarkably, the effective Hamiltonian [Eq. (117)] for such a truncated version of quantum electrodynamics (the so-called Jaynes–Cummings model) has a number of important properties, and experimental studies by Haroche and Walther and others were begun in the early 1980s. Exact expressions are known for the eigenvalues and eigenvectors of H_{JC}. With $\hbar = 1$ for simplicity and $\Delta = \omega_{21} - \omega$, the eigenvalues are

$$E_{n,\pm} = E_1 + n\omega + \tfrac{1}{2}[\Delta \pm \Omega_n], \tag{118a}$$

where we have reinserted $E_1 \neq 0$ for the lower level energy, and the corresponding eigenvectors are given by

$$\|n,+\rangle = \cos\Phi|n-1\rangle * |+\rangle + \sin\Phi|n\rangle * |-\rangle \tag{118b}$$

$$\|n,-\rangle = -\sin\Phi|n-1\rangle * |+\rangle + \cos\Phi|n\rangle * |-\rangle, \tag{118c}$$

where $\cos\Phi = \sqrt{(\Omega_n + \Delta)/2\Omega_n}$ and $\sin\Phi = \sqrt{(\Omega_n - \Delta)/2\Omega_n}$. Here the use of Ω_n for $\sqrt{4\lambda^2 n + \delta^2}$ is a deliberate reminder of Ω defined following Eqs. (12) because they play the same roles in their respective quantum and semiclassical theories. Similarly, one writes $\chi(n)$ as a reminder of χ for the same reason, that is, $\chi(n) = 2\lambda\sqrt{n}$ is the QED equivalent of the Rabi frequency $2d\mathscr{E}_0/\hbar$ [and not to be confused with the χ_m of Eqs. (49) and (50)].

One of the observable quantities of the Jaynes–Cummings model is the atomic energy. Its expectation value can be calculated exactly, without approximation:

$$\langle\sigma_z(t)\rangle = -\sum_n p_n \cos\chi(n)\,t \quad \text{(quantum).} \tag{119a}$$

Here p_n is the probability that the single mode has exactly n photons. This result can be contrasted with Eq. (18c) in

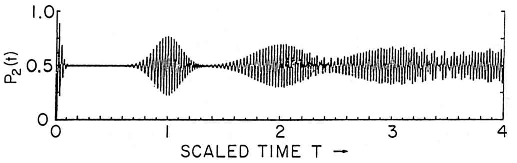


FIGURE 10 Quantum collapse and revival of atomic inversion.

the same limit $\Delta = 0$ and under the same circumstances, namely, with the field intensity $I \approx \chi^2$ distributed with some probability $p(I)$ over a range of values:

$$w(t) = -\int p(I)\cos\chi(I)t\,dI \quad \text{(semiclassical).} \tag{119b}$$

The apparent similarity of these results disguises fundamental differences in dynamical behavior between Eqs. (119a and b). These differences arise from the discreteness of the allowed photon numbers in Eq. (119a), which are discrete precisely because the field is quantized, that is, because the mode contains only whole photons and never fractional units of the energy $\hbar\omega$. By contrast, in the semiclassical theory (recall Section I.C) the intensity I and Rabi frequency χ can have any values.

For the quantum photon distribution associated with a coherent state, for which p_n is given by Eq. (90), a plot of $\langle\sigma_z(t)\rangle$ is shown in Fig. 10. The nearly immediate disappearance "collapse" of the signal shortly after $t = 0$ can be explained on the basis of the interference of many frequencies $\chi(n)$ in the quantum sum of Eq. (119a). Such a collapse can be expected for any broad distribution p_n and would be predicted by the semiclassical Eq. (119b) as well, if $p(I)$ is a broad distribution function. However, the predicted reappearances or "revivals" of the signal are a sign that the field is quantized. They occur, and at regular intervals, only because p_n is a discrete distribution. The semiclassical expression (119b) leads inevitably to an irreversible collapse. Only quantum theory can provide the stepwise discontinuous photon number distribution that is the basis for the revivals.

The revivals and other quantum mechanical predictions implied by the truncated Hamiltonian [Eq. (117)] are of interest because this Hamiltonian is simple enough to permit exact calculations, without further approximations of the kind familiar in most of radiation theory. For example, the expression for quantum inversion in Eq. (119a) has the following unusual properties:

1. it is not restricted to any finite range of t values;
2. it holds for all values of the coupling constant λ, which is contained in $\chi(n) = 2\lambda\sqrt{n}$;
3. it is completely free of decorrelations, such as the commonly used approximation $\langle a^\dagger a\sigma_z\rangle \approx \langle a^\dagger a\rangle\langle\sigma_z\rangle$;
4. it is finite even at exact resonance ($\omega_{21} = \omega$) without the aid of ad hoc complex energies;
5. it is fully quantum mechanical with nontrivial commutators preserved: $[a, a^\dagger] = 1$, $[\sigma, \sigma^\dagger] = 2\sigma_z$, etc; and
6. it is realistically nonlinear (it saturates because the atomic energy cannot exceed E_2).

This combination of properties is unique in atomic radiation theory. They indicate, for example, that a system obeying Hamiltonian [Eq. (117)] would permit some fundamental questions in quantum electrodynamics to be studied independently of the restrictions of the usual perturbation methods that are based on short-time expansions and a small coupling constant. Experimental realization of the model is unlikely in the optical frequency range because of the restriction to a single radiation mode. However, quantum optical techniques, including the detection of single photons, are rapidly being extended to much lower frequencies, where single-mode cavities can be built. Observation of the Jaynes–Cummings model is expected to play a guiding role in microwave single-mode experiments with Rydberg atoms.

E. AC Stark Effect and Resonance Fluorescence

Just as the Bloch vector provides a powerful descriptive framework for a wide variety of quantum optical phenomena, so does the Jaynes–Cummings model. The Hamiltonian [Eq. (117)] can be regarded as a zero-order approximation to a "true" Hamiltonian, in which the atom is allowed more than two levels or the field has more than one mode. It is an unsual zero-order approximation because it includes the interacting Hamiltonian as well as the noninteracting atomic and radiation Hamiltonians.

If the atom interacts resonantly with a single strong mode of the field, then its interactions with other modes, perhaps involving other levels of the atom, can be treated approximately. The energy spectrum of the Jaynes–Cummings Hamiltonian makes it clear how this can be done. In Fig. 11 the RWA energy spectrum is shown in the absence of a strong resonant interaction (i.e., with $\lambda = 0$), and also with $\lambda \neq 0$. The spectrum shows that the state $|n\rangle * |1\rangle$, which corresponds to the atom in its lower level and n photons in the mode when $\lambda = 0$, is pushed down to become the state $\|n, -\rangle$ when $\lambda \neq 0$. That is, for Eq. (118a) we obtain

$$E_{n,-} = E_1 + n\omega - \tfrac{1}{2}[\Omega_n - \Delta]. \tag{120}$$

Since $\Omega_n \geqq \Delta$, the lower level is pushed down. Similarly, the corresponding upper state $|n\rangle * |2\rangle$ is pushed up by the same amount. This shift is called the *AC Stark shift* because it is due to the interaction with an oscillating (alternating) electric field. The size of the AC Stark shift δ_{AC} varies as a function of Δ in the range

$$\tfrac{1}{2}\chi_n \geq \delta_{\mathrm{AC}} \geq \chi_n^2/4\Delta \tag{121}$$

depending on whether the atom and the field mode are near to or far from resonance.

An external probe of the coupled two-level plus single-mode system can reveal these details of its spectrum. For example, the two nearly degenerate states $\|n, +\rangle$ and $\|n, -\rangle$ are split by twice the AC Stark shift. This splitting can be observed by absorption spectroscopy if a weak second radiation field is allowed to induce transitions to a third level in the atom. This was first described and observed in 1955 by Autler and Townes.

A different kind of probe is provided by *resonance fluorescence*, that is, by spontaneous emission into modes other than the main mode. In this case a strong laser field provides the main mode radiation. Dipole selection rules determine that from the same nearly degenerate states $\|n, +\rangle$ and $\|n, -\rangle$ spontaneous transitions can be made only to the next lower pair of nearly degenerate states, $\|n-1, +\rangle$ and $\|n-1, -\rangle$. There are four separate emission lines predicted, as shown in Fig. 11. The strengths of the four lines are all equal on resonance, but since two of them have the same transition frequency only three lines are actually expected. They have the intensity ratio 1:2:1, but the side peaks have different widths than the center peak, and the peak height ratio is 1:3:1. This fluorescence triplet was predicted in the late 1960s, and after a period of controversy about the exact line structure, the predictions mentioned here were verified experimentally in the mid-1970s by Stroud, Walther, Ezekiel, and others. It should be clear from Fig. 11 that the resonance fluorescence peak separation is equal to the Autler–Townes splitting, which is just the quantum Rabi frequency $\chi(n)$ at resonance. The sequence of spectra shown in Fig. 12 illustrates the increased peak separation that accompanies an increased Rabi frequency when the main mode intensity is increased.

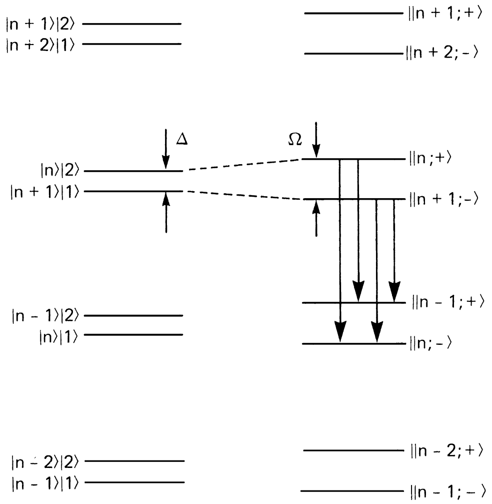


FIGURE 11 Jaynes–Cummings RWA energy spectrum.

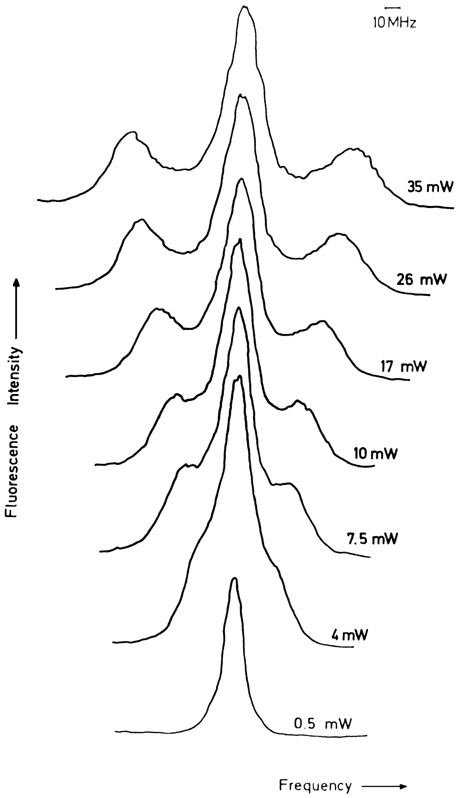


FIGURE 12 Resonance fluorescence spectra in the presence of AC Stark splitting. [Reprinted with permission from Hartig, W., Rasmussen, W., Schieder, R., and Walther, H. (1976). Study of the frequency distribution of the fluorescent light induced by monochromatic radiation. *Zeitschrift für Physik* **A278,** 205–210.]

F. Tests of Quantum Theory

It has long been recognized that quantum theory stands in conflict with naive notions that "physical reality" can be independent of observation. As is well known, the *Heisenberg uncertainty principle* mandates a limit on the mutual precision with which two noncommuting variables may be observed. This curious feature was put into sharp focus by Einstein, Podolsky, and Rosen in 1935, but for nearly half a century it remained mainly of philosophical interest, as experimental tests were difficult to conceive or implement. The situation changed dramatically during the 1970s when it was realized that quantum optical experiments could test different conceptions of "reality."

Einstein, Podolsky, and Rosen (EPR) gave a precise meaning to the concept of reality in this context: "If, without in any way disturbing a system, we can predict with certainty [i.e., with probability equal to unity] the value of a physical quantity, then there exists an element of physical reality corresponding to this physical quantity." It was of primary concern to EPR whether quantum theory can be considered to be a "complete" theory. A necessary condition for completeness of a theory, according to EPR, is that "every element of the physical reality must have a counterpart in the physical theory." Using these definitions of reality and completeness, and the properties of correlated quantum states, EPR concluded that quantum theory does not provide a complete description of physical reality.

An illuminating example due to Bohm is provided by the singlet state of two spin$-\frac{1}{2}$ particles:

$$\|S\rangle = (1/\sqrt{2})[|a+, \hat{\mathbf{n}}\rangle * |b-, \hat{\mathbf{n}}\rangle - |a-, \hat{\mathbf{n}}\rangle * b+, \hat{\mathbf{n}}\rangle], \tag{122}$$

where $|a\pm, \hat{\mathbf{n}}\rangle$ is the state for which particle a has spin up $(+)$ or down $(-)$ in the direction $\hat{\mathbf{n}}$. The unit vector $\hat{\mathbf{n}}$ can point in any direction. In light of the EPR argument, one notes that the spin of particle b in, say, the $\hat{\mathbf{x}}$ direction, can be predicted with certainty from a measurement of the spin of particle a in that direction. If the spin of particle a is found to be up, then the spin of particle b must, for the singlet state, be down, and vice versa. Thus, the spin of particle b in the $\hat{\mathbf{x}}$ direction can be predicted with certainty "without in any way disturbing" that particle. According to EPR, therefore, the $\hat{\mathbf{x}}$ component of the spin of particle b is an element of physical reality.

Of course, one may choose instead to measure the $\hat{\mathbf{y}}$ component of the spin of particle a, in which case the $\hat{\mathbf{y}}$ component of particle b can be predicted with certainty. It follows therefore that both the $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ components of the spin of particle b (and, of course, particle a) are elements of physical reality. However, according to quantum mechanics, the $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ components of spin cannot have simultaneously predetermined values, because the associated spin operators do not commute. Therefore, quantum theory does not account for these elements of physical reality, and so, according to EPR, it is not a complete theory.

The EPR experiment can be criticized on the ground that "the system" must be understood in its totality and cannot refer to just one of the particles, for instance, of a correlated two-particle system. In the Bohm gedanken experiment just considered, a measurement on particle a does disturb the complete (two-particle) system because the quantum mechanical description of the system is changed by the measurement. One cannot consistently associate an element of physical reality with each spin component of particle b, even though the particles may be arbitrarily far apart and not interacting in any way.

Motivated by the EPR argument, one may ask whether it is possible to formulate a theory in which physical quantities do have objectively real values "out there," independently of whether any measurements are made. These objectively real values may be imagined to be determined by certain *hidden variables* which may themselves be stochastic. One can ask, is it possible, in principle, to construct a hidden variable theory in full agreement with the statistical predictions of quantum mechanics, but which allows for an objective reality in the EPR sense?

In the early 1960s Bell considered the most palatable class of hidden variable theories, the so-called local theories. He demonstrated that such theories cannot fully agree with quantum mechanics. In particular, certain *Bell inequalities* distinguish any local hidden variable theory from quantum mechanics. Another way of stating "Bell's theorem" is that no local realistic theory can be in full agreement with the predictions of quantum theory.

Bell inequalities are not difficult to derive for the Bohm thought experiment. A local hidden variable theory is postulated to give $\pm\frac{1}{2}$ for each spin component, as determined by the hidden variables. The theory is also supposed to account for the spin correlations: if one is up the other must be down. The difference between such a theory and quantum mechanics is that the spins are predetermined (by the hidden variables) before any measurement, that is, they are objectively real. The condition of locality enters through the additional assumption that a measurement of the spin of each particle is not affected by the direction in which the spin of the other particle is measured. This is certainly reasonable if all the spin components are predetermined, since the two particles may be very far apart when a measurement is made.

The question now is whether any measurements can distinguish such a theory from quantum mechanics. Bell considered $E(\hat{\mathbf{m}}, \hat{\mathbf{n}})$, the expectation value of the product of the spin components of particles a and b in the $\hat{\mathbf{m}}$ and $\hat{\mathbf{n}}$ directions, respectively. He obtained the inequality

$$|E(\hat{\mathbf{m}}, \hat{\mathbf{n}}) - E(\hat{\mathbf{m}}, \hat{\mathbf{p}})| \leqq \tfrac{1}{4} + E(\hat{\mathbf{n}}, \hat{\mathbf{p}}), \tag{123}$$

which must be satisfied by the entire class of local hidden variable theories. This inequality is violated by quantum theory, as can be seen from the quantum mechanical prediction $E(\hat{\mathbf{m}}, \hat{n}) = -\frac{1}{4}\hat{\mathbf{m}} \cdot \hat{n}$. It is therefore possible to test experimentally the predictions of quantum theory vis-à-vis the whole class of plausible "realistic" theories.

Although Bell's theorem promoted philosophical questions about hidden variables to the level of experimental verifiability, it remained difficult to conceive of specific experiments that could be undertaken. In 1969, however, Clauser and co-workers suggested that Bell inequalities could be tested by measuring photon polarization correlations if certain additional but reasonable assumptions about the measurement process were made. The spin considered by Bell is replaced by photon polarization, another two-state phenomenon. Correlated two-photon polarization states are produced in atomic cascade emissions, and efficient polarizers and detectors are available for optical photons.

Consider a $J = 0 \to 1 \to 0$ atomic cascade decay, with polarizer–detector systems on the $\pm z$ axes. Linear polarization filters may be employed to distinguish the photons by their energy so that each polarizer–detector system records photons of one frequency but not the other. It may be shown that the two-photon prolarization state has the form

$$\|\Psi\rangle = (1/\sqrt{2}[|\mathrm{a}, \hat{\mathbf{x}}\rangle * |\mathrm{b}, \hat{\mathbf{y}}\rangle + |\mathrm{a}, \hat{\mathbf{y}}\rangle * |\mathrm{b}, \hat{\mathbf{x}}\rangle]), \tag{124}$$

where $|\mathrm{a}, \hat{\mathbf{x}}\rangle$ is the single-photon state in which photon a is linearly polarized along the $\hat{\mathbf{x}}$ direction, etc. A similar form applies if a circular polarization basis is used.

The correlated photon state of Eq. (124) is obviously analogous to the spin-$\frac{1}{2}$ correlated state of Eq. (122). A hidden variable theory of such polarization correlations leads to a Bell inequality analogous to Eq. (123):

$$|E(\alpha, \beta) - E(\alpha, \gamma)| \leq 1 - E(\beta, \gamma), \tag{125}$$

where $E(\alpha, \beta)$ now refers to photon polarization components and α and β are the filter orientations with respect to some reference axis. The differences between Eqs. (125) and (123) arise because we are now dealing with spin-one particles and because Eq. (124) describes a positive correlation. The quantum mechanical prediction for $E(\alpha, \beta)$ is simply cos $2(\alpha - \beta)$, and Eq. (125) becomes

$$|\cos 2(\alpha - \beta) - \cos 2(\alpha - \gamma)| \leqq 1 - \cos 2(\beta - \gamma), \tag{126}$$

which, in fact, is not satisfied for all angles α, β, and γ. Such violations of Bell inequalities have been observed in independent experiments led by Clauser, Fry, and Aspect. The results of such experiments are in agreement with quantum theory, and appear to rule out any local hidden variable theory. There are possible loopholes in the interpretation of the experiments, but at the present time none of them seem very plausible. According to Clauser and Shimony, “The conclusions are philosophically unsettling: Either one most totally abandon the realistic philosophy of most working scientists, or dramatically revise our present concept of space–time.”

From the viewpoint of quantum optics, the photon polarization correlation experiments measure a second-order field correlation function. Such a correlation function not only distinguishes between classical and quantum radiation theories, but also between quantum theory and local realistic theories. In quantum optics it is often possible to address such questions from essentially first principles and to carry out accurate tests of theory in the laboratory.

V. RECENT DEVELOPMENTS

A. Substantial Squeezing

In Section III.B the first observation of a squeezed state of light was mentioned. Since then, considerably greater degrees of squeezing have been reported. As much as a 60% noise reduction has been observed using a three-wave parametric down-conversion technique, in which a photon of frequency ω is converted into two photons of frequency $\omega/2$ in a crystal.

B. Two-Photon Correlated Quantum States

Studies of photon coherence and interference have entered a new domain with experiments being undertaken on new types of two-photon quantum states, going beyond the detection of two-photon polarization correlations mentioned in Section IV.F. For example, quantum beats have been observed in the joint probability of two-photon detection as a function of the path difference and frequency difference of two photons produced by parametric down conversion.

C. Cavity QED in the Optical Domain

Despite our prediction to the contrary at the end of Section IV.D, experimental observations of effects associated with the single-atom single-mode Jaynes–Cummings quantum model have been successful in the optical domain, as well as at microwave frequencies. Collapse and revival signals (see Fig. 10) have been reported in experiments using rubidium atoms traversing a microwave cavity, and optical observations of line-narrowing below the free-space limit and cavity line-spilitting have been reported.

D. Two-Photon Laser

After more than two decades of theoretical interest, substantial experimental progress toward the realization of a two-photon laser is occurring. A two-photon laser has many fundamental differences with the conventional one-photon laser, including special noise properties, multistability, and a different phase transition at threshold. An important step has been the successful operation of a quantum oscillator working on a microwave two-photon Rydberg transition in rubidium.

E. Strong-Field Quantum Optics and Atomic Continuum States

According to the considerations of Section II.F, ionizing transitions should be immune to saturation. Since the continuum of states above the ionization threshold is infinitely broad ($\beta \to \infty$), no finite χ should satisfy Eq. (47). This has turned out not to be true. New laser systems which deliver 0.01–1.0 terawatts (1 terawatt $= 10^{12}$ W) in 1 ps (or shorter) pulses have been used in multiphoton atomic ionization experiments that have revealed new phenomena for which saturation and other quantum optical concepts seen unexpectedly appropriate. These phenomena include equally spaced multiple-peaked photoelectron spectra and anomalously strong very high-order harmonic generation (above 30th order).

F. Cooling below the Doppler Limit

Resonant excitation of two-level atoms can give rise to center of mass motion as well as internal transitions, and various schemes for trapping and cooling collections of atoms by laser light are based on this. The spontaneous line-width Γ of the transition places a so-called Doppler limit $kT = \hbar\Gamma/2$ on the lowest temperature T that can be reached. However, it has been realized that multilevel atoms allow this limit to be evaded, and much lower temperatures, in the neighborhood of 50 μK (0.000050 degrees absolute) have now been recorded.

G. Teleportation

One key consequence of the violation of the Bell inequality is to establish the nonlocality of quantum mechanics. Modern quantum optical techniques permit nonlocality to be exploited when one "teleports" a quantum state, and teleportation has been demonstrated in several laboratories. Teleportation here means to send an ideally exact copy of a quantum state localized at the sender (familiarly called Alice) to a remote site (where the operator is known as Bob). Teleportation is not in conflict with the "no cloning" theorem because the original state to be teleported is lost in the process, so twins are never created.

We will discuss teleportation of the polarization state of a photon labeled 1 in Alice's control. We write the photon state as:

$$|\Psi_1\rangle = a|H_1\rangle + b|V_1\rangle, \quad \text{where } |a|^2 + |b|^2 = 1,$$

where H and V refer to horizontal and vertical polarization, respectively (or any pair of crossed polarization directions). Two more photons, labeled 2 and 3, are used to teleport this state. They need to be arranged in a Bell-correlated state, for example

$$\|\Psi_{23}\rangle = (1/\sqrt{2})(|H_2, V_3\rangle - |V_2, H_3\rangle),$$

as described in Section IV.F. Photons 2 and 3 are "shared" in the sense that 2 is directed to Alice and 3 to Bob, but Bob and Alice know only that they are sharing a Bell pair and nothing about its specific nature. Thus this illustration of teleportation starts with a three-photon quantum system, whose state $\||\Psi_{123}\rangle$ is the tensor product of the state to be teleported and the Bell state: $\||\Psi_{123} \equiv |\Psi_1\rangle * \|\Psi_{23}\rangle$.

The process of teleportation is easily unraveled in the *Bell basis* of entangled two-photon states. The Bell basis for the product space of particles 1 and 2 is the following four states labeled A, B, C, D:

$$\left\|\Psi_{12}^{(A)}\right\rangle = \tfrac{1}{\sqrt{2}}(|H_1\rangle|H_2\rangle + |V_1\rangle|V_2\rangle),$$

$$\left\|\Psi_{12}^{(B)}\right\rangle = \tfrac{1}{\sqrt{2}}(|H_1\rangle|H_2\rangle - |V_1\rangle|V_2\rangle),$$

$$\left\|\Psi_{12}^{(C)}\right\rangle = \tfrac{1}{\sqrt{2}}(|H_1\rangle|V_2\rangle + |V_1\rangle|H_2\rangle),$$

$$\left\|\Psi_{12}^{(D)}\right\rangle = \tfrac{1}{\sqrt{2}}(|H_1\rangle|V_2\rangle - |V_1\rangle|H_2\rangle).$$

It is not hard to check that our three-photon state $\||\Psi_{123}\rangle$, can be written:

$$\begin{aligned}\||\Psi_{123}\rangle = {} & \tfrac{1}{2}\left\|\Psi_{12}^{(A)}\right\rangle * (-b|H_3\rangle + a|V_3\rangle)\\ & + \tfrac{1}{2}\left\|\Psi_{12}^{(B)}\right\rangle * (b|H_3\rangle + a|V_3\rangle)\\ & + \tfrac{1}{2}\left\|\Psi_{12}^{(C)}\right\rangle * (-a|H_3\rangle + b|V_3\rangle)\\ & + \tfrac{1}{2}\left\|\Psi_{12}^{(D)}\right\rangle * (-a|H_3\rangle - b|V_3\rangle).\end{aligned}$$

Alice starts the process by performing what is called a *Bell measurement* on the two photons available to her (1 and 2). In doing this Alice collapses the three-photon jointly held state $\||\Psi_{123}\rangle$ to one of the Bell components. Next Alice calls Bob on the telephone to tell him which component. This signals Bob to make not a measurement but a unitary transformation on the state of the photon he has received, photon 3. The transformation he makes is determined universally by the following teleportation table:

Alice's Bell-state projection	**Bob's unitary operation needed**
$\|\Psi_{12}\rangle \to \left\|\Psi_{12}^{(A)}\right\rangle$	$i\sigma_y = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$
$\|\Psi_{12}\rangle \to \left\|\Psi_{12}^{(B)}\right\rangle$	$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$
$\|\Psi_{12}\rangle \to \left\|\Psi_{12}^{(C)}\right\rangle$	$-\sigma_z = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$
$\|\Psi_{12}\rangle \to \left\|\Psi_{12}^{(D)}\right\rangle$	$-\hat{1} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$

This table is constructed so that, after Bob performs the indicated unitary transformation, particle 3 is in the state

$|\Psi_3\rangle = a|H_1\rangle + b|V_1\rangle$. Note that this is the same as the original state of particle 1, and the teleportation is complete. The reader can check this by choosing the state 3 partner of any one of the Bell components in $\||\Psi_{123}\rangle$ and performing the rotation indicated by the table.

At this point Bob has the original state while Alice has the entangled Bell pair resulting from her Bell projection, but nothing of her original state of photon 1. The original state is destroyed in the process of teleporting it. Alice's inability to know in advance which entangled Bell state will result from her Bell measurement is mirrored in Bob's uncertainty about the unitary operation (polarization basis rotation) required of him before receiving Alice's phone call. This uncertainty necessitates the phone call, and this eliminates any imagined superluminal aspect of the process. The locations of Alice and Bob are irrelevent to the process and they may be separated by a great distance.

SEE ALSO THE FOLLOWING ARTICLES

Atomic and Molecular Collisions • Electromagnetics • Lasers • Nonlinear Optical Processes • Optical Interferometry • Quantum Theory

BIBLIOGRAPHY

Allen, L., and Eberly, J. H. (1975). "Optical Resonance and Two-Level Atoms," Wiley, New York, reprinted by Dover (1987).

Bennett, C. H., Brassard, G., Crepeau, C., Jozsa, R., Peres, A., and Wooters, W. K. (1993). *Phys. Rev. Lett.* **70,** 1895.

Boschi, D., Branca, S., De Martini, F., Hardy, L., and Popescu, S. (1997). *Phys. Rev. Lett.* **80,** 1121.

Bouwmeester, B., Pan, J. W., Mattle, K., Eibl, M., Weinfurter, H., and Zeilinger, A. (1997). *Nature* **390,** 575.

Delone, N. B., and Krainov, V. P. (1985). "Atoms in Strong Light Fields." Springer-Verlag, Berlin.

Fontana, P. (1982). "Atomic Radiative Processes," Academic Press, San Diego.

Furusawa, A., Sorenson, J., Braunstein, S. L., Fuchs, C. A., Kimble, H. J., and Polzik, E. S. (1998). *Science* **282,** 706.

Knight, P. L., and Allen, L. (1983). "Concepts of Quantum Optics," Pergamon, Oxford.

Knight, P. L., and Milonni, P. W. (1980). The Rabi frequency in optical spectra. *In* "Physics Reports," Vol. 66, pp. 21–107. North-Holland, Amsterdam.

Loudon, R. (1983). "The Quantum Theory of Light," 2nd ed. Oxford Univ. Press, Oxford, UK.

Mandel, L. (1976). The case for and against semiclassical radiation theory. *In* "Progress in Optics" (E. Wolf, ed.), Vol. XIII. North-Holland, Amsterdam.

Milonni, P. W. (1976). Semiclassical and quantum-electrodynamical approaches in nonrelativistic radiation theory. *In* "Physics Reports," Vol. 25, pp. 1–81. North-Holland, Amsterdam.

Perina, J. (1984). "Quantum Statistics of Linear and Nonlinear Phenomena," D. Reidel, Dordrecht.

Rosen, H. J., and Gustafson, T. K. (eds.) (1989). Quantum electronic applications and optical studies of high-T_c superconductors. *Quantum Electron* **25**(11), 2357–2409.

Stenholm, S. (1984). "Foundations of Laser Spectroscopy," Wiley, New York.

Yoo, H. I., and Eberly, J. H. (1985). Dynamical theory of an atom with two or three levels interacting with quantized cavity fields. *In* "Physics Reports," Vol. 118, pp. 239–337. North-Holland, Amsterdam.

Quantum Theory

David W. Cohen
Smith College

GLOSSARY

Continuous Pertaining to a variable that can assume all values from a specified continuum.

Continuum Interval of real numbers, possibly unbounded.

Discrete Pertaining to a set of numbers in which every member is isolated from the others in the set by an open interval. Also pertaining to a variable that can assume only values from a specified discrete set of numbers.

Macroscopic Pertaining to physical phenomena occurring on a scale large enough to be observed without the aid of a microscope.

Quantum Quantity of energy.

A QUANTUM THEORY is a system of explanations of physical phenomena based on the assumption that certain quantities can assume only a discrete set of numerical values. The first formulation of a quantum theory was presented in 1900 by Max Planck to explain results of experiments that measured light radiating from a heated body. Planck broke a tradition of classical physics by assuming that energy was a discrete variable in one situation in which it had always been considered continuous. He introduced two other ideas that also may be considered hallmarks of a quantum theory. One was the hypothesis that discrete quantities of energy (energy quanta) can be treated as objects, like particles, subject to statistical laws of distribution—a foreshadowing of the merger of wave theory and particle theory. The second was the introduction of a new constant to the general laws of physics.

Between 1900 and 1925 there emerged the quantum theories of blackbody radiation, light, specific heat, and atomic energy, among others. This is the body of physics commonly called the quantum theory. Beginning in 1925 all of these theories, as well as the formulation of the philosophical questions behind them, were merged into one mathematical formalism called quantum mechanics.

Our discussion of the quantum theory will trace the main ideas in chronological order from the classical setting in the 19th century through the birth of quantum mechanics in 1925. Then we will look at some of the controversy surrounding the modern theoretical foundations of quantum mechanics. And finally we will look at an example of how the probabilistic view of nature arising from quantum theory might actually be used to advantage to build high-speed quantum computers. In our discussions we won't hesitate to paraphrase original ideas and substitute modern notation when it's necessary for clear understanding.

I. THE CLASSICAL SETTING

In this section we review the classical physics that we shall need for our discussion of quantum theory.

A. Physical Laws and Variables

One of Sir Isaac Newton's key contributions to science in the 17th century was the concept of the physical law. He codified events, like the fall of an apple, into a system of words and mathematical expressions that brought such order, clarity, and economy to thought that he was able to reveal hidden relationships among the events. In fact, Newton's laws provided experimenters with the means to predict with uncanny certainty how some events would affect the future.

The success of the predictive value of Newton's laws was so great that it became a universally accepted principle of science that it is possible, in theory, to determine simultaneously the values of all physical quantities associated with a given physical event and with those values predict, with 100% certainty, all future events caused by the first. We shall see later how quantum theory challenged this principle, but first let us examine the law of Newtonian mechanics that we shall need for our discussion. We use modern terminology.

Let us call a physical quantity fundamental if its values are determined by comparing measurements of the quantity against some arbitrarily selected standard. Examples of such quantities are periods of time (measured in seconds, perhaps) and length (measured in meters, feet, etc.). We shall call derived quantities those obtained from fundamental quantities by mathematical calculations. Examples of these are instantaneous velocity and acceleration.

One fundamental quantity defined by Newton is the mass of an object. Mass measures the resistance of objects to acceleration. An object used to determine a standard value of mass, 1 kg, is kept in a vault in Paris. We can define a derived quantity, force, by setting particles of known mass into motion. When an object of mass m is accelerated from rest, along a straight line, to an acceleration a, we say the force F causing the acceleration acts in the direction of the motion and is given by

$$F = ma. \tag{1}$$

Notice that force, like acceleration, is described in terms of direction, and so it is a vector quantity. The magnitude of force necessary to accelerate 1 kg of mass to a magnitude of 1 m sec^{-2} is called 1 newton (N). Equation (1) is a formulation of Newton's second principle of motion.

Let us look more closely at the meaning of Eq. (1). We think of the letters in the equation as place holders for numbers. That is, if we assign numbers to two of the letters, we can manipulate the equation to compute the number that must be assigned to the third letter, if the equation is to be true. So we call the letters variables. A variable in an equation is called continuous if the numbers that can be assigned to it comprise an open (possibly unbounded) interval. For example, if we assume that we can accelerate an object of given mass to any magnitude of acceleration between values $-x$ and x, then the variable a in Eq. (1) is continuous with domain of assignability $(-x, x)$.

Some variables, however, can be assigned values only from a discrete set of numbers. A set of numbers S is called discrete if every number in S belongs to an open interval that contains no other members of S. The set of integers, for example, is a discrete set of numbers. In the equation $\cos 2N\pi = 1$, the set of values for N for which the equation is true can come only from the set of integers. Thus, N is a discrete variable in that equation.

The word *discrete*, as we have defined it, is relatively modern. Physicists traditionally have used the word *discontinuous* instead. One of the revolutionary ideas of quantum theory was the assumption of the discontinuity of some variables that had always been considered continuous in classical physics.

B. Mechanical Energy

Energy is the capacity to do work. We present here, briefly, the classical definitions of potential and kinetic energy for a particle moving in a field of force.

Let R denote a region of three-dimensional space and suppose that at every point p in R there is a three-dimensional vector $F(p)$ assigned to that point. We call F a vector field. Suppose also that f is a real-valued function on R, that ∇f is its gradient, and that

$$F(p) = -\nabla f(p) \tag{2}$$

for every point p of R. Any function f that satisfies Eq. (2) is called a potential function for F. If $F(p)$ represents a force at every point p, we call F a force field.

If γ is any differentiable path in R from point $p=\gamma(a)$ to point $q=\gamma(b)$, and if f is a potential function for force field F, then we define the work done by moving a particle along γ through force field F by the line integral

$$W=\int_{\gamma} F. \tag{3}$$

The work depends only on the end points of the path, and not on γ itself. We can express this by the equations

$$\int_{\gamma} F=\int_{a}^{b} F(\gamma(t))\cdot\gamma'(t)\,dt=f(q)-f(p), \tag{4}$$

which can be proved using Eq. (2).

We have, then, that $f(q)-f(p)$ is the work done if we move a particle through force field F from point p to point q. Since any two potential functions for F must differ only by a constant, that amount of work can be calculated from Eq. (4) using any potential function for F. We call that amount of work the potential energy difference between point p and point q.

It is possible to define an absolute potential energy function PE by selecting any point s in R and deciding what the potential energy P is to be at the point. Then for any potential function f for F, we can define the potential energy at every point p to be

$$\mathrm{PE}(p)=P+f(p)-f(s). \tag{5}$$

This function PE differs from f by a constant, so it is a potential function for F, and its definition does not depend on the choice for f.

An example of a potential energy function arises when a particle moves about R, all of three-dimensional space, in a field of a force of attraction between the particle and the point O at the origin of R. The potential energy can be negative at every point p, with large negative values near O and the potential energy approaching zero at points very far from O. Then to move the particle from a distant point p along a straight line toward O to a point s closer to O requires negative work, because it is movement in the same direction as the force exerted by the field. This results in a change of potential energy from one negative value to a more negative value, which is considered a decrease in potential energy. As we shall see, an electron in an atom is subject to a potential energy function of this type.

Now let us consider kinetic energy. A force field F that satisfies Eq. (2) for some potential function f is called a conservative force field. A particle placed in a conservative force field and allowed to react solely to the force exerted by the field will move from point to point, always in the direction of decreasing potential energy. During the motion, we ascribe to the particle at each point p a kinetic energy defined by

$$\mathrm{KE}(p)=mv(p)^2/2, \tag{6}$$

where m is the mass of the particle and $v(p)$ the magnitude of the velocity of the particle when it is at point p. The kinetic energy is equal to the amount of work that must be done to accelerate the particle from a state of rest to a state of motion with velocity of magnitude $v(p)$.

The law of conservation of energy states that, throughout the motion of the particle, the total energy

$$E(p)=\mathrm{PE}(p)+\mathrm{KE}(p) \tag{7}$$

has the same value for every point p.

A useful example is that of the kinetic energy of a particle of mass m rotating at constant speed in a circular orbit of radius r. Suppose that at every point the particle is subject to a force toward the center of the orbit of magnitude F and that its velocity at every point has magnitude v. It follows from some routine calculations that at each point the particle is subject to an acceleration toward the center of the orbit of magnitude $a=v^2/r$. Then we have from Eq. (1) that $F=ma=mv^2/r$, so that we can obtain the kinetic energy of the particle at every point in its orbit as

$$\mathrm{KE}=\tfrac{1}{2}mv^2=\tfrac{1}{2}Fr. \tag{8}$$

Later, we shall make use of this equation.

C. Electromagnetic Energy

In the late 19th century James Clerk Maxwell combined theories of electricity and magnetism into a theory of electromagnetic fields. Such a field is a region of space at every point of which is a vector of electromagnetic force that acts on any charged particle placed at the point. The work done as the force moves the particle is a measure of the electromagnetic energy of the field. Such fields surround every charged particle, and if a charged particle is accelerated, some of the field surrounding it leaves the vicinity of the particle and travels off into space. We call this traveling field a wave of electromagnetic radiation. We can measure the energy of the radiation by measuring the work the wave can perform in moving a charged particle placed in its path.

It will not be necessary for our discussion of quantum theory to examine the quantitative relationships involved with electromagnetic theory, but it is important to know that these relationships, known as Maxwell's equations, require fields of electromagnetic force to be continuous vector fields. Electromagnetic energy calculated from Maxwell's equations is necessarily a continuous variable.

D. Entropy

Entropy was originally defined as a measure of the relationship between temperature and energy. The total mechanical energy due to the motion of the molecules in a given substance is called the thermal energy of the substance. Some of this thermal energy can be made to do work if the substance is allowed to cool.

In the early part of the 19th century Sadi Carnot provided a theoretical model of an engine that transforms a fall in temperature of a substance into mechanical energy. Some years later Rudolf Clausius and William Thomson (Lord Kelvin) observed the following about Carnot's engine: Of the total difference in thermal energy resulting from a drop in temperature from T_1 (degrees Kelvin) to T_2 in a Carnot engine, only the fractional portion $1 - T_2/T_1$ can be transformed into mechanical energy to do work, even under the most ideal theoretical conditions. The remaining portion of energy is redistributed in the substance of the engine. Note that no work can be obtained from thermal energy without a strict decrease in temperature.

The principle of Carnot's engine applies to a gas at temperature T_1 occupying a volume V_1 and exerting pressure P_1 on the walls of its container. Let us say that this gas is in a macroscopic state $w_1 = (P_1, V_1, T_1)$. Suppose we wish to change the state of the gas to $w_2 = (P_2, V_2, T_2)$. We could, if we wished, change states without changing temperature by allowing the gas to expand slowly against one movable side of its container, decreasing pressure and increasing volume. Since work, or mechanical energy, is obtained from this process (assuming the side resists movement), we can add thermal energy to the gas to counteract the temperature drop necessary to supply the work. On the other hand, we could, at least theoretically, allow the gas to expand in a vacuum, increasing volume without changing pressure. In this case, no work is done and temperature can remain constant without an addition of thermal energy. The first type of change from w_1 to w_2, requiring an addition (or subtraction) of thermal energy, if temperature is to remain constant, is called a reversible change of state.

Now suppose we wish to pass in tiny reversible incremental steps $w_1, w_2, \ldots, w_n$ from some initial state w_1 to some final state w_n. Denote by ΔE_k the change in thermal energy for the gas as it goes from state w_k to w_{k+1}. Now consider the number

$$S(w_1, w_n) = \sum_{k=1}^{n} \frac{\Delta E_k}{T_k}. \tag{9}$$

Clausius proved that for reversible changes of state, the sum of Eq. (9) depends only on w_1 and w_n and is independent of the sequence used in changing from the initial to the final state. He called this intrinsic difference between the two states a difference in the entropy of the gas. Notice that, as with potential energy, we do not have a value for the absolute entropy until we assign a value to one state.

The theory of entropy was eventually extended to physical systems other than fixed amounts of gas. A physical system can consist of a mixture of gases or particles of different kinds, or a region of space containing items of unknown nature. Even the entire universe can be considered a giant physical system. Extending the theory of entropy was not easy. Among the most difficult tasks was assigning meaning to the notion of the "state" of various physical systems. We shall return to that problem later, but let us continue our discussion of entropy supposing that "states" have been defined.

As in the theory of gases, changes in the state of a physical system can be classified as reversible or nonreversible. If we consider a sequence $w_1, w_2, \ldots, w_n$ of reversible changes in state for a system at absolute temperature T, as we did to obtain Eq. (9), then the incremental change in entropy from state w_k to w_{k+1} is denoted

$$\Delta S_k = \Delta E_k / T, \tag{10}$$

where ΔE_k is the incremental change in the total energy of the system.

We examine our next step carefully. It reveals an assumption taken so much for granted in the physics of the 19th century that it was seldom, if ever, explicitly discussed then. The challenge to that assumption was at the very heart of quantum theory in the early 20th century.

We have in Eq. (10) a relationship between differences in entropy and differences in energy. For a gas in a controlled experiment, precise values for these differences can be measured. If we assign a particular value of entropy to a particular value of energy in an experiment, it is possible to use Eq. (10) to write an explicit relationship between absolute entropy and energy at temperature T for states $w_1, w_2, \ldots, w_n$:

$$S_k = E_k / T. \tag{11}$$

Next comes the assumption. We assume that the states $w_1, w_2, \ldots, w_n$ were chosen from a continuum of states so that Eq. (11) provides a discrete set of values in what is really a continuous, in fact, differentiable function of entropy in terms of energy. Time and again physicists identified one physical quantity as a function of another one, and assumed that the functional relationship was continuous and that the ability to measure arbitrarily small differences in the quantities was limited only by the ability to build arbitrarily precise and accurate measuring devices. Although this assumption will be reexamined in our discussion of quantum theory, it allows us to take our next step to obtain an expression involving entropy that we shall need in the sequel.

We rewrite Eq. (10) as $\Delta S_k / \Delta E_k = 1/T$ and use our assumption to pass to infinitesimally small incremental changes to obtain an expression for the derivative of entropy with respect to energy:

$$dS/dE = 1/T. \tag{12}$$

This equation was a basic building block in the original construction of the quantum theory.

E. Specific Heat

As we mentioned in the preceding section, the thermal energy of a substance is transformed into mechanical energy when the substance undergoes a drop in temperature. Conversely, energy supplied to a substance can raise its temperature. Let us consider cases in which a substance does not change pressure (if it is a gas) or volume while it undergoes changes in temperature.

Throughout the 19th century experiments were performed to determine the amount of energy required to raise fixed amounts of various substances one degree on a temperature scale. It was discovered that the relation between the amount of energy added to a substance and the resulting rise in temperature depends not only on the material of the substance but also on the temperature of the material during the addition of energy.

Let us consider a fixed substance. Consider a sequence $T_1, \ldots, T_n$ of temperatures. Denote by ΔT_k the change in temperature from T_k to T_{k+1}, and by ΔE_k the corresponding change in thermal energy of the substance. Now we define the heat capacity of the substance at temperature T_k as

$$C(T_k) = \Delta E_k / \Delta T_k. \qquad (13)$$

Using the assumption of differentiability mentioned in our discussion of entropy, we can then pass to infinitesimally small increments in Eq. (13) to arrive at the definition for the heat capacity of the substance at temperature T as

$$C(T) = dE(T)/dT. \qquad (14)$$

The specific heat of a substance at temperature T is defined as its heat capacity per unit mass. Thus, a substance of mass m has specific heat at temperature T given by

$$c(T) = C(T)/m. \qquad (15)$$

Although we have defined heat capacity and specific heat as functions of temperature, one often hears a phrase such as “the specific heat of copper” without mention of temperature. That is because the values of the heat capacity of most substances are nearly constant over a range of room temperatures, and it is this constant value that is often called the heat capacity of the substance.

F. Boltzmann Statistics

In the late 19th century Ludwig Boltzmann undertook the task of providing a more precise relationship between changes in mechanical motions of molecules and changes in entropy. His theory rested on notions of distributing numbers according to laws of probability, and this notion was a key inspiration to the developers of quantum theory.

Let us consider a gas in a closed container and the macroscopic states of the gas defined as $w = (P, V, T)$ when the gas is at pressure P, volume V, and temperature T. We shall restrict our attention to changes in macroscopic state due solely to changes in temperature, leaving pressure and volume fixed. Consider one macroscopic state w, and suppose that the gas consists of N molecules, each with a specific energy, and that in state w the total thermal energy of the gas is E.

Next we shall partition the real-number interval $[0, E]$ into p subintervals $J_1, \ldots, J_p$, each of length $\varepsilon = E/p$. We then assign each of the N molecules to one subinterval according to the amount of energy it has. For a given assignment we consider a p-tuple of numbers $(l_1, \ldots, l_p)$, where l_i is the number of molecules assigned to subinterval J_i. We call such a p-tuple a microscopic state of the gas. For example, if $l_1 = N$ and $l_2 = l_3 = \cdots = l_p = 0$, then the gas is in microscopic state $(N, 0, \ldots, 0)$, and we can write the total energy of the gas as $E = p\varepsilon \approx N\varepsilon = l_1\varepsilon$. The approximation follows since all N molecules have energy values in interval J_1, which has the number ε as an end point. So each molecule has energy approximately equal to ε. For an arbitrary microscopic state $(l_1, \ldots, l_p)$, we have an approximation of E given by

$$E \approx \sum_{i=1}^{p} l_i x_i, \qquad (16)$$

where each x_i is an arbitrary value in subinterval J_i.

It is assumed that the molecules are distinguishable, so that there are many ways a specific microscopic state can be achieved. In other words, interchanging one of the molecules in subinterval J_m with one in subinterval J_n does not change the microscopic state, which depends only on the p-tuple $(l_1, \ldots, l_p)$. It can be computed from standard methods in combinatorics that the total number of ways the molecules can be assigned among the subintervals to achieve a given microscopic state $(l_1, \ldots, l_p)$ is

$$W(l_1, \ldots, l_p) = N!/l_1!l_2! \cdots l_p! \qquad (17)$$

Now if $W(l_1, \ldots, l_p)$ is a high number, many assignments of the molecules to the subintervals result in state $(l_1, \ldots, l_p)$. If it is a small number, few assignments result in that state. The common way to express this in physics is to say that $W(l_1, \ldots, l_p)$ is the probability for microscopic state $(l_1, \ldots, l_p)$ to occur. A state where $W(l_1, \ldots, l_p)$ achieves a maximum value is a state “most likely to occur,” and that is called a state of statistical thermal equilibrium.

Another way to describe W is to call it a measure of the “disorder” of the molecules. States highly likely to occur are those having coordinates nearly equal. They are said to be in a state of high disorder. That is why it is often said that entropy is a measure of the disorder of the universe.

Now if we have two systems that we combine into one (e.g., two gases in the same vessel), then an increase in entropy S_1 and S_2 of the systems individually must result in an increase $S_1 + S_2$ of the entropy of the combined system. On the other hand, if W_1 is the probability for a given state for one system with entropy S_1 to occur, and W_2 is the probability for a given state for the other system with entropy S_2 to occur, then the probability for the combined system to be in both states simultaneously is the product $W_1 W_2$. So the relationship between entropy S and the probability W of a microscopic state of a system must be the one that relates multiplication to addition,

$$S = k \ln W, \tag{18}$$

where k is a constant. This is Boltzmann's key entropy equation, although he never wrote it this way. The constant k, now known as Boltzmann's constant, was the subject of much research at the beginning of the 20th century.

Observe that, at thermal statistical equilibrium for a given energy, W is a maximum and hence S is also. A macroscopic state of statistical thermal equilibrium, therefore, is a state of maximum entropy for a given energy, and that macroscopic state can occur for many different microscopic states.

Let us conclude our introduction to Boltzmann statistics by mentioning a step that Boltzmann took whenever he applied his calculations. By using the assumption of continuity we discussed in Section I.D, he let the number of subintervals of the energy interval go to infinity and the size of each subinterval go to zero. As we shall see, the success of quantum theory rested on *not* taking that step.

II. BLACKBODY RADIATION: THE BIRTH OF QUANTUM THEORY

A. The Blackbody Furnace

Particles absorb and radiate electromagnetic energy. Some absorb very little incident radiation (reflectors), and others a great deal (absorbers). Objects can be made to radiate electromagnetic energy when they are heated. We see this when we watch glowing coals in a dying fire, for light is a form of electromagnetic radiation. The most intense energy is radiated by objects that are the best absorbers. A perfect absorber is called a blackbody.

Physicists can simulate a blackbody by building a box and placing a tiny hole in one side of the box. The hole is the blackbody, because nearly all electromagnetic energy falling on the hole will enter into the box and not emerge again. We shall call such a box a blackbody furnace.

Suppose we build such a furnace and heat the interior of it and hold it at constant temperature T (degrees Kelvin). The particles in the furnace, dust, for example, and the particles making up the walls of the furnace will absorb and radiate electromagnetic energy until the interior reaches a state of equilibrium. That is when each particle is absorbing exactly the same amount of energy it is radiating. In reality, if a tiny bit of radiation energy escapes through the hole, more energy is being emitted than absorbed by some particles, but we can ignore that tiny difference and consider the electromagnetic energy in the furnace to be a field of electromagnetic energy in a state of equilibrium.

We now wish to measure the intensity per unit volume of the electromagnetic energy in the furnace. In practice we can do this by letting a small amount of electromagnetic energy radiate from the box through the hole. When this was done in the late 19th century, it was found that the intensity varied with the temperature T and the frequency ν of the radiation. This can be observed, for example, by shining the light from the furnace through a prism and measuring the intensity of the light at different frequencies. Other variables were considered. For a given temperature and frequency the size and the shape of the furnace were varied. Furnaces were built from a variety of materials, and particles of different materials were placed in the furnaces. None of these changes affected the intensity of the electromagnetic energy, which appeared to depend *only* on frequency and temperature. Thus, it was hypothesized that there exists an expression I for the energy intensity in a unit volume in the furnace as a function of frequency ν and temperature T that is fundamental to blackbody radiation.

In the late 19th century there began an intense and frustrating search for the function I based on a sound theoretical model and verified by experiment. Well-known laws of classical electromagnetic radiation, thermodynamics, and mechanics were blended together to give formulations for I that were patently absurd. On the other hand, expressions for I extrapolated from experimental data were accurate to various extents but stood wholly without theoretical foundations. The search for an explanation of blackbody radiation led to the birth of quantum theory.

B. The Extrapolated Formulas

Let us begin the search for the function I in the same way it was begun by physicists around the turn of the century: by examining experimental data. For various temperatures $T_1, \ldots, T_n$, we plot the observed values for $I(\nu, T_k)$ as a function of ν. Four such plots are illustrated in Fig. 1.

For a specific temperature T, if $I(\nu_1, T)$ and $I(\nu_2, T)$ are the energy intensity per unit volume at frequencies ν_1 and ν_2, respectively, then the total energy intensity in

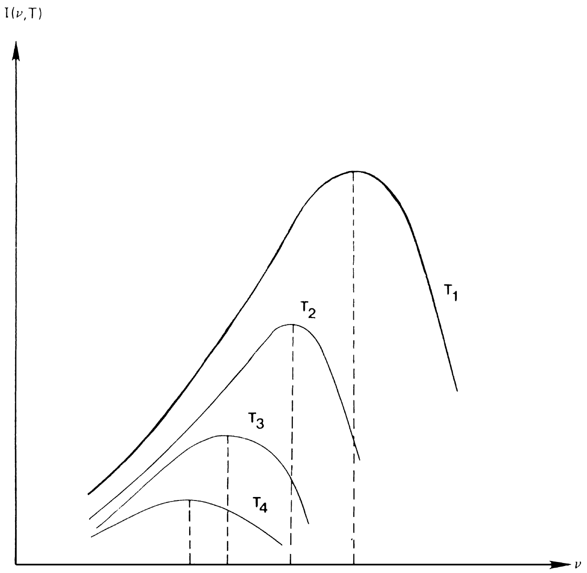


FIGURE 1 Graphs of energy intensity per unit volume as a function of frequency for blackbody radiation in thermal equilibrium at four different temperatures.

a unit volume containing electromagnetic energy of all frequencies between ν_1 and ν_2 is

$$E(T, \nu_1, \nu_2) = \int_{\nu_1}^{\nu_2} I(\nu, T)\, d\nu. \tag{19}$$

Because the total energy over a range of frequencies (called a spectral range) is given by an integral, I is called a spectral density function. If all frequencies are present, the total energy in a unit volume is

$$E(T) = \int_0^{\infty} I(\nu, T)\, d\nu, \tag{20}$$

a function of temperature only. It was an observation of Josef Stefan in 1879 that $E(T)$ was proportional to the fourth power of T. This observation, as well as others gathered from empirical data, became guideposts in the search for a general theory of blackbody radiation, for no theory can gain wide acceptance if it cannot be shown to be in agreement with available experimental evidence.

Another observation we can make from Fig. 1 is that for any given temperature there is a single frequency of maximum intensity, and this frequency depends on the temperature. This characteristic of $I(\nu, T)$ is known as the displacement property, because an increase in temperature "displaces" the frequency of maximum intensity farther to the right on the ν axis. In 1894 Wilhelm Wien set down a displacement formula that was verified experimentally by Friedrich Paschen in 1899.

Wien also published an expression for $I(\nu, T)$ in 1896, which agreed with much of the experimental data available at the time. His formula was in the form

$$I(\nu, T) = a\nu^3 \exp(-b\nu/T), \tag{21}$$

where a and b are constants. Soon after it was published, Wien's formula was found to be inaccurate for low values of ν/T.

Between 1897 and 1900 Max Planck worked on obtaining a better formula, one with a theoretical explanation based on classical principles of physics. No satisfactory explanation had accompanied Wien's empirical formula [Eq. (21)], and as we shall see in the next section, it was Planck's search for this explanation that led him to quantum theory.

C. Planck's Empirical Law

A scientific theory sometimes seems to one who reads about it to be a stroke of genius, and often it is. What the reader does not know, however, is that often the author had a peek at the "right answer" before developing the theory. Such was the case with Max Planck, who presented his theory of blackbody radiation to the Physikalische Gesellschaft in Berlin on December 14, 1900. That is the date often cited as the birthday of quantum theory. We begin our examination of Planck's theory in the same way he developed it, by peeking at the "right" radiation law, which he himself had found.

Let us suppose that the walls of the blackbody furnace and the particles in the cavity are all composed of tiny linear harmonic oscillators (Planck originally called them resonators) that emit and absorb electromagnetic energy. Using classical laws of thermodynamics and electromagnetism, it can be shown that

$$I\,(\nu, T) = a\nu^2 E(\nu, T), \tag{22}$$

where a is a constant and $E(\nu, T)$ is the average (over time) of the energy intensity of a typical oscillator vibrating at frequency ν. It would take us too far afield to follow the derivation of Eq. (22), but it is important to note here that the classical laws of electromagnetic energy on which it relies require that the amplitude and energy of the vibrations of a linear oscillator be continuously variable.

If we accept this model of linear oscillators, the search for I becomes a search for E to fit into Eq. (22). Planck used the second law of thermodynamics and a keen intuition for interpolation between formulas known to be successful in different portions of the temperature scale to arrive at an expression for E. The result, Planck's radiation law, is of the form

$$E(\nu, T) = h\nu/[\exp(B\nu/T) - 1], \tag{23}$$

where B and h are constants. This formula when placed into Eq. (22), gives excellent agreement with experiment at a wide range of temperatures and frequencies. Convinced that he had the "right" radiation formula, Planck then set about finding the precise mechanical model for the furnace and a theory about how it operated to yield this result.

D. The Quantum Theory of Blackbody Radiation

If our basic assumption is that the furnace and the particles in it are composed of linear oscillators, two questions arise immediately: How many oscillators are there, and how are the total radiant energy and total entropy of the furnace distributed among the oscillators?

We shall answer the first question, as did Planck, in the simplest way, by assuming that there are finitely many oscillators for each frequency ν, say N_ν. If we write $S(\nu, T)$ for the entropy of a linear oscillator in thermal equilibrium at temperature T and frequency ν, and $E(\nu, T)$ for its average energy, then we have for the total entropy and total energy of all oscillators having frequency ν

$$S_{\text{tot},\nu} = N_\nu S(\nu, T) \quad \text{and} \quad E_{\text{tot},\nu} = N_\nu E(\nu, T). \tag{24}$$

Our search for E continues on the basis of a relation between E and S.

If we solve Eq. (23) for T and use Eq. (12), which relates energy to entropy, by writing $dS(\nu, T)/dE(\nu, T) = 1/T$, we can perform an integration to obtain the entropy for a single oscillator,

$$S(\nu, T) = \frac{h}{B} \ln \left[\frac{[1 + E(\nu, T)/h\nu]^{1+E(\nu,T)/h\nu}}{[E(\nu, T)/h\nu]^{E(\nu,T)/h\nu}} \right] + D, \tag{25}$$

where D is a constant of integration. We shall return to this equation later.

The next step was described by Planck as "an act of desperation." He had been a dedicated opponent of Boltzmann's view that a total amount of energy in a gas was distributed among a finite number of molecules according to laws of probability and combinations. Laws of probability permit exceptions, and Planck thought that such laws had no place in a theory of thermodynamics, which he held to be valid without exceptions. Yet now he saw that such a view point could be taken to move him along toward a derivation of his own radiation law. With a flexibility that often distinguishes the creative genius from the pedant, Planck adopted ideas of Boltzmann that he had previously rejected. Let us continue following Planck's reasoning.

To obtain discrete quantities of energy to distribute over our oscillators, we divide the total energy $E_{\text{tot},\nu}$ into p elements of size ε, so that

$$E_{\text{tot},\nu} = p\varepsilon = N_\nu E(\nu, T). \tag{26}$$

To simplify notation we shall now drop the symbols ν and T in what follows, bearing in mind that we have N oscillators, all at equilibrium at temperature T and vibrating with frequency ν.

According to Boltzmann's theory, if a system is in a given microscopic state with probability W, then the total entropy S is proportional to ln W,

$$S = k \ln W, \tag{27}$$

where k is a constant. The task now is to calculate W for our system of oscillators. This will be the total number of ways the p energy elements can be distributed among the N oscillators. Note that, although we are taking a cue from Boltzmann, our approach differs in that our "states" reflect distributions of p numbers among N oscillators, whereas Boltzmann's states reflect distributions of N molecules over p intervals of real numbers.

Let us count the possible distributions of energy elements among oscillators. Suppose, first, that one oscillator is assigned all p energy elements. There are N ways that this distribution could occur. That is each of the N oscillators in turn could be assigned all the energy. Another distribution might be to assign all but one energy unit to one oscillator and the remaining one to another. There are $N(N-1)$ ways to achieve this distribution. It is a routine calculation in combinatorics to show that the total number of ways to achieve all possible distributions is

$$W = (N + p - 1)!/p!(N - 1)! \tag{28}$$

Since N is very large for practical purposes, one can neglect the subtraction of 1 in Eq. (28) to obtain

$$W = (N + p)!/p!N! \tag{29}$$

We simplify W further by using Stirling's formula $x! \approx x^x$ to obtain

$$W = (N + p)^{N+p}/N^N p^p. \tag{30}$$

We now have two expressions for the total entropy of the system. After an algebraic manipulation using Eq. (25), we obtain the total entropy (for frequency ν and temperature T):

$$\begin{aligned} S_{\text{tot}} &= NS \\ &= \frac{Nh}{B}\left[\left(1 + \frac{E}{h\nu}\right)\ln\left(1 + \frac{E}{h\nu}\right) \right. \\ &\qquad \left. - \frac{E}{h\nu}\ln\left(\frac{E}{h\nu}\right)\right] + D. \end{aligned} \tag{31}$$

At the same time we have directly from Eq. (27), using Eq. (30), a bit of algebraic manipulation, and the substitution $p/N = E/\varepsilon$ from Eq. (26):

$$\begin{aligned} S_{\text{tot}} &= k \ln W \\ &= k[(N+p)\ln(N+p) - N\ln N - p\ln p] \\ &= kN\left[\left(1+\frac{p}{N}\right)\ln\left(\left(1+\frac{p}{N}\right)N\right)\right. \\ &\quad \left. - \ln N - \frac{p}{N}\ln p\right] \\ &= kN\left[\left(1+\frac{E}{\varepsilon}\right)\ln\left(1+\frac{E}{\varepsilon}\right) - \frac{E}{\varepsilon}\ln\frac{E}{\varepsilon}\right]. \qquad (32) \end{aligned}$$

Both Eqs. (31) and (32) give the total entropy contributed by N oscillators. Dividing each by N gives two expressions for the entropy S of a single oscillator in terms of energy E and frequency ν, still at temperature T. How can we make both expressions for S equivalent? The answer is clear. Set $h/B = k$ and set

$$\varepsilon = h\nu. \qquad (33)$$

This last equation provides the key that now allows us to construct a theory to *derive* Planck's radiation formula (23). This is the quantum theory.

We begin with the supposition that the furnace is composed of linear oscillators and that, for a fixed frequency, the total entropy contributed by N oscillators in thermal equilibrium with that frequency is given by Eq. (27), where W is interpreted according to our discussion following that equation. We then follow the manipulation of Eq. (27), which we used to derive Eq. (32). Setting $\varepsilon = h\nu$, designating ε as a "quantum" of energy, and then dividing Eq. (32) by N, we obtain the entropy S for a single oscillator in our theory. If we differentiate this expression for S with respect to E, and set the derivative equal to $1/T$ according to Eq. (12), we then can solve for E. This derivation yields the "correct" radiation law:

$$E(\nu, T) = h\nu/[\exp(h\nu/kT) - 1]. \qquad (34)$$

Thus, the quantum theory was born.

When Planck put this formula into Eq. (22) and then integrated over ν to get $E(T)$, he obtained Stephan's fourth power law and verified Wien's displacement law, each of which he wrote in terms of the new constant h. Knowing the experimental values for these laws, he was able to compute a value for h:

$$h = 6.55 \times 10^{-27} \text{ erg sec (energy times time).} \qquad (35)$$

This constant was to emerge at the very heart of many theories of physics throughout the 20th century and, in fact, was to play a crucial role in the fundamentals of quantum mechanics.

III. EARLY EVOLUTION

A. Difficulties with Planck's Theory

Some of the difficulties with Planck's theory of blackbody radiation were obvious immediately; others were quite subtle and were not discovered until several years after the presentation of the theory in 1900.

Consider, first, the fuzzy relationship between Planck's use of probabilism and Boltzmann's. As we mentioned in Section I.F, after Boltzmann partitioned the energy interval into finitely many subintervals, he later allowed the number of subintervals to approach infinity and the size of each subinterval to approach zero. That was required to apply the classical continuity assumptions he needed in applications of his theory. Planck carefully noted in his address to the Physikalische Gesellschaft in 1900 that his energy quanta $\varepsilon = h\nu$ must not be allowed to tend toward zero. The finiteness of the number of oscillators N_ν makes it essential to maintain the finiteness of the number of energy quanta in order to apply the combinatoric procedure associated with Eq. (27).

There was another discrepancy between Planck's theory and Boltzmann's statistical mechanics. It had been a generally accepted principle of statistical mechanics that, in an aggregate of oscillators in thermal equilibrium, all with the same number of degrees of freedom, the toal energy of each oscillator must, on average (over time), be distributed equally among its degrees of freedom. This principle was a consequence of what was called the equipartition theorem. If Planck had applied the theorem to his oscillators, then instead of Eq. (34) he would have obtained $E(\nu, T) = kT$ and would have arrived at an incorrect radiation law. Planck's theory violated the principle of equipartition. It is not completely clear whether Planck was even aware of this principle in 1900.

A more fundamental difficulty, a logical inconsistency, was recognized by Albert Einstein in 1905. Planck had originally thought of his partitioning of the total energy into discrete quantities as a mathematical device to obtain numbers to treat with probabilistic arguments. He did not realize until it was pointed out by Einstein that, for his derivation to be consistent, each of his oscillators had to be assumed to be able to absorb and emit energy only over a discrete range of values. On the other hand, Planck's derivation of Eq. (22) requires that the oscillators be able to absorb and emit energy over a continuum of values. It is therefore inconsistent to put Eq. (32) together with Eq. (22) to arrive at a radiation law.

Despite the difficulties, Planck's theory of radiation was acknowledged for the accuracy of the formula resulting from it, and history shows that the theory itself revolutionized physics. The "discontinuity" (more accurately, the "discreteness") of the energy variable and the statistical nature of the behavior of discrete energy quanta were ideas that were to become the foundations of a new and controversial view of the universe.

Albert Einstein, of course, was as important to quantum theory as he was to nearly every other development of physics in the early 20th century. Sometimes a friend and sometimes a foe of the rapidly evolving quantum theory, he made important contributions to it and, merely by paying attention to it, helped to spur the interest of the scientific community. Let us now discuss two ideas of Einstein that were instrumental in placing the "quantum" in the forefront of physics.

B. Einstein's Theory of Light Quanta

Einstein set forth his theory of light in one of three papers published in 1905. He won the Nobel prize for that paper, and although it is commonly referred to as his paper on the "photoelectric effect," it really was much more.

Einstein was disturbed by the dualism in physics between particle mechanics and the electromagnetic wave theory of Maxwell. The fundamental difference was the discrete nature of the former as opposed to the continuous nature of the latter. The continuous wave theory of light was inadequate to explain some experimental phenomena. For example, it was known that ultraviolet light incident on a piece of metal causes electrons to be emitted from the metal. Contrary to electromagnetic wave theory, however, the velocity of an emitted electron does not depend on the intensity (wave energy determined by amplitude) of the incident light, but instead is a function of its frequency. This phenomenon is known as the photoelectric effect.

One of Einstein's motives for investigating light was to explain this effect, although his bold explanation was such a violent departure from accepted theory of the wave nature of light that it had implications far beyond consideration of the photoelectric effect. In particular, it had a profound influence on the development of quantum theory, although, ironically, Planck rejected Einstein's theory of light.

We begin our development of the theory of light by returning to light radiation in the cavity of a blackbody furnace in thermal equilibrium at temperature T. As before, consider the spectral density function $I(\nu, T)$, which gives the energy intensity per unit volume in the cavity due to the portion of radiation having frequency ν.

Let us write $S(I, \nu)$ for an entropy density function, which gives the entropy per unit volume as a function of energy intensity and frequency ν, and consider that our first task is to find the correct expression for $S(I, \nu)$ in terms of I and ν. Our notation is a bit redundant, since I is a function of ν, but it will make our subsequent calculations easier to follow.

We shall follow Einstein and use Wien's expression for I,

$$I(\nu, T) = a\nu^3 \exp(-b\nu/T) \tag{21}$$

instead of Planck's, even though it was well known that Wien's law was valid only for large values of ν/T. Solving Eq. (21) for $1/T$, we obtain

$$1/T = -\ln[I(\nu, T)/a\nu^3]/b\nu. \tag{36}$$

Then, using the fact that the derivative of entropy with respect to energy is equal to $1/T$, we get

$$\partial S(I, \nu)/\partial I = 1/T. \tag{37}$$

We then substitute Eq. (36) into Eq. (37) and integrate with respect to I to obtain the entropy function

$$S(I, \nu) = \frac{-I(\nu, T)[\ln(I(\nu, T)/a\nu^3) - 1]}{b\nu}. \tag{38}$$

Now consider a volume V in the cavity, and suppose that the radiation is nearly monochromatic, say of frequency ν. The radiation energy in volume V for frequency ν is given by another spectral density function,

$$E(\nu, T) = VI(\nu, T). \tag{39}$$

The entropy in volume V for frequency ν is then found by solving Eq. (39) for $I(\nu, T)$ and substituting into Eq. (38):

$$\begin{aligned} S(E, \nu) &= VS(I, \nu) \\ &= \frac{-E(\nu, T)[\ln(E(\nu, T)/Va\nu^3) - 1]}{b\nu}. \end{aligned} \tag{40}$$

Next we recalculate a new entropy for a volume V' smaller than V. For the same energy $E(\nu, T)$, substitute V' into Eq. (40) to calculate S'. Then the entropy decreases by an amount

$$S(E, \nu) - S'(E, \nu) = \frac{-E(\nu, T)\ln(V'/V)}{b\nu}. \tag{41}$$

Now we relate Eq. (41) to the key relation of Boltzmann statistics, $S = k \ln W$, which applies to an ideal gas in a closed container. (Einstein used this equation in his paper of 1905 but did not express it using the letter k.) For a change of the gas from one state W to a state W', the corresponding change in entropy is

$$S - S' = -k \ln(W'/W), \tag{42}$$

where W is interpreted as the "probability" or likelihood for the given state to occur.

Boltzmann statistics applies to a finite number of molecules in a gas. We are not considering a gas in blackbody radiation, nor are we explicitly using Planck's model of finitely many "oscillators." What then shall we consider as the meaning of a "state" in our context? Returning to our derivation of Eq. (42) in terms of a change in volume from V to V', let us recall that, if N particles (of anything) are allowed to travel freely in a vessel of volume V, then the probability of finding, at any given instant, all N particles in a portion of the vessel having volume $V' < V$ is

$$(V'/V)^N. \tag{43}$$

The corresponding decrease in entropy of the system of particles from the state of random distribution to the state of all particles in the smaller portion is thus

$$\begin{aligned} S - S' &= -k \ln(V'/V)^N \\ &= -Nk \ln(V'/V). \end{aligned} \tag{44}$$

Now we argue backward from Eqs. (41) and (44). If we take the view that Eq. (41) gives the change in entropy accompanying a decrease in the volume of the blackbody cavity from V to V', and we assume that the radiation energy $E(\nu, T)$ is distributed among N independent particles of some sort, which are allowed to move freely in the cavity, then the right side of Eq. (41) must equal the right side of Eq. (45). We then conclude that

$$Nk = E(\nu, T)/b\nu \quad \text{or} \quad E(\nu, T) = Nkb\nu. \tag{45}$$

That is, monochromatic radiant energy behaves as if it were composed of independent energy "quanta" of magnitude $kb\nu$.

Einstein immediately suggested that the same reasoning could be applied to the radiation of light and thus fired another shot in the 20th century revolution in physics by proposing a return to the particle theory of light: a view considered dead nearly a century. This proposal was so radical that even Planck, who never considered himself a revolutionary anyway, rejected it.

Let us see how Einstein's theory of light quanta provided an explanation of the photoelectric effect. If an electron in a metal strip is set free by the energy of a quantum of light incident on the metal, then the kinetic energy of the electron is equal to at most the energy of the incident quantum, which is proportional to the frequency of the light. This kinetic energy is less than the energy of the incident light quantum, according to the amount of work required to overcome the forces tending to keep the electron bound to the metal. Moreover, increasing light intensity increases the number of light quanta incident on the metal and hence increases the number of electrons freed; but the energy of the freed electrons is dependent solely on the frequency of the light.

Einstein was careful to point out that his light quantum hypothesis, which was motivated in part by Wien's radiation law, was limited to the range of frequencies and temperatures for which Wien's law was valid.

Einstein's approach to quantization differed from Planck's in a fundamental way. Planck assumed the quantization of energy to derive his radiation law, therefore showing that the quantum hypothesis was *sufficient* to obtain the law. Einstein, on the other hand, started with Wien's radiation law and showed that one of its *necessary* consequences was the quantization of monochromatic radiant energy.

It is also worth mentioning the advantage Einstein's theory had over Planck's in that it did not involve oscillators and so did not directly contradict the equipartition theorem.

Let us now look briefly at the data predicted by Einstein's theory and verified to a high degree of accuracy by R. A. Millikan in 1916, about 10 years after Einstein published his paper. We shall denote the maximum kinetic energy of an electron emitted by a light quantum of frequency ν as

$$\mathrm{KE} = kb\nu - E', \tag{46}$$

where E' is the amount of energy necessary to remove the electron from the metal. Note that this kinetic energy is a linear function of ν. When accurate date are plotted as a linear graph of KE versus ν, we can measure the slope of the line, which turns out to be Planck's constant h. Since the nature of the metal affects the kinetic energy of the electron only in the additive constant E', the slope h appears to be a universal constant. In other words, $kb = h$ in Eq. (46), so that the energy of a light quantum of frequency ν is $h\nu$.

Although Einstein and Planck were at odds over the fundamentals of each other's work, this constant h provided an undeniable link between their theories and a powerful motivating force for further investigations into its role in nature.

C. The Quantum Theory of Specific Heat

At the same time Einstein put forth his theory of light he also proposed a theory of specific heat of solids that was to become another pillar of quantum theory. His work further widened the range of applicability of quantum concepts.

Let us return to Planck's radiation law,

$$E(\nu, T) = h\nu/[\exp(h\nu/kT) - 1] \tag{34}$$

for the average (over time) energy of an oscillator of frequency ν in a blackbody furnace in equilibrium at temperature T. Recall that a consequence of Planck's derivation of his radiation law is that the oscillators in the radiation field can have energy values only in the discrete range $0, h\nu, 2h\nu, \ldots,$ for these are the values of the energy elements he uses to distribute over the oscillators to make his statistical calculations. Einstein understood that this implied a revolutionary principle that in turn implied a modification of all theories of physical phenomena dealing with exchanges of energy between radiation and matter. In particular, the theory of heat, based on the equipartition law, was at odds with experimental evidence on the specific heat of solids. He sought a new theory of specific heat without the equipartition law, similar to the way Planck had disregarded equipartition in his theory leading to Eq. (34).

Although Einstein used Eq. (34) in his theory of specific heat, it is important to note that he provided his own derivation of it. To get at the heart of Einstein's derivation, let us consider a vibrating physical system (e.g., a linear oscillator) that can exist in various states of thermodynamic equilibrium, while vibrating at frequency ν. Let us denote by $\phi(E, \nu)$ the energy density function, which, when integrated between limits E_1 and E_2, gives the number of states that have energy between those limits. Let P be a probability density function: $P(E, T)$ is the probability that the system is in a state of energy E when at equilibrium at temperature T. Then the energy of the system is given by the expected value

$$E(\nu, T) = \frac{\int_0^\infty EP(E, T)\phi(E, \nu)\, dE}{\int_0^\infty P(E, T)\phi(E, \nu)\, dE}. \qquad (47)$$

We now arrive at Einstein's characterization of ϕ, which defines a principle applicable to all systems involving interaction between vibrating matter and electromagnetic radiation.

There exists a positive number r, very small compared with $h\nu$ (h is Planck's constant), and a sequence of intervals of numbers $I_n = [nh\nu, nh\nu + r]$ such that:

1. $\phi(E, v) \neq 0$ only for values of E lying in the intervals $I_0, I_1, \ldots.$.
2. $\int_{I_n} \phi\,(E, \nu)\, dE$ is equal to the same constant for all integers $n = 0, 1, 2, \ldots.$

Figure 2 shows intervals I_n for $n = 0, 1, 2, 3$.

From this formulation of ϕ and the use of expression $P(E) = \exp(-E/kT)$ (from statistical mechanics) in Eq. (47), Einstein was able to derive Eq. (34).

What is accomplished by this formulation of ϕ is the restriction of nonzero energy values to a set of small intervals: again, almost a "quantization" of energy. Much more is also accomplished, however. First, the value $h\nu$ emerges as a key energy value with an existence not immediately tied to linear oscillators. Second, as we shall see later, this generalized derivation of Eq. (34) enabled Peter Debye to use it as a point of departure to obtain Planck's radiation law without the fundamental inconsistencies contained in Planck's derivation. Finally, the formulation of ϕ as a principle applying to all systems involving interaction between matter and radiation elevated the status of quantum theory from an ad hoc assumption about oscillators to a scientific principle.

Let us apply Eq. (34) to vibrating atoms, which determine the heat capacity of a solid. If an atom has three independent vibrational degrees of freedom, all vibrating with the same frequency ν, then it can be considered to be three independent vibrating systems, and so the energy of each atom is calculated from Eq. (34):

$$E(\nu, T) = 3h\nu/[\exp(h\nu/kT) - 1]. \qquad (48)$$

The energy of 1 g-atom of the solid is therefore

$$E(\nu, T) = 3Nh\nu/[\exp(h\nu/kT) - 1], \qquad (49)$$

where N, known as Avogadro's number, is the number of atoms in 1 g-atom of the solid in question. From Eq. (14), the heat capacity of the solid is then

$$\frac{dE_{\text{tot}}(\nu, T)}{dT} = \frac{3Nk(h\nu/kT)^2 \exp(h\nu/kT)}{[\exp(h\nu/kT) - 1]^2}. \qquad (50)$$

The formula for the specific heat of the solid then follows by dividing Eq. (50) by the mass of the solid. This result was recognized by Einstein as subject to correction because it rested on serveral simplification assumptions.

In spite of its need for correction, Eq. (50) represents the first application of quantum theory to solids. What is more important, historically, is that it established that the principle of quantum theory reached beyond one or two specific physical phenomena. This was the view of Walther Nernst, who obtained solid evidence of Einstein's formula for specific heat and whose eloquent praise of quantum theory around 1910 was instrumental in the organization of the first Solvay Congress in 1911. That was a conference of distinguished physicists gathered together in Brussels for the purpose of discussing the "new" quantum theory.

D. Debye's Derivation of Planck's Radiation Law

As we have noted, Planck's theory leading to his radiation law contains a fundamental inconsistency made clear by Einstein: the requirement that oscillator energy by

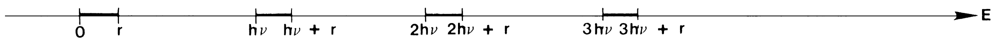


FIGURE 2 Intervals of nonzero values of energy density $\phi(E, \nu)$ in Einstein's derivation of the law of blackbody radiation.

continuously variable for the formula $I(\nu, T) = a\nu^2 E(\nu, T)$ and that it be discontinuously variable for the formula $E(\nu, T) = h\nu/[\exp(h\nu/kT) - 1]$. Nevertheless, the unqualified success of the radiation formula itself in matching experimental results, together with the successful application of the fundamental idea of quantization to other realms of physics, prompted great efforts to remove the inconsistency. This was finally accomplished by Peter Debye in 1910. Let us trace Debye's argument.

The key to our success will be the elimination altogether of the need for oscillators. Instead, we consider the cavity of a blackbody furnace to be a resonating chamber and draw an analogy between the radiation waves in the electromagnetic field in the cavity and the vibrations of an elastic fluid in an enclosed container. This idea was, in fact, proposed in 1900 by J. W. Strutt, better known as Lord Rayleigh, who was an expert in sound and likened blackbody radiation to sound waves.

It is known from the theory of sound that a cubic box of volume L^3 supports standing sound waves of vibration only in modes of wavelength

$$\lambda = 2L/\sqrt{k^2 + l^2 + m^2}, \tag{51}$$

where k, l, and m are positive integers.

Let R^3 stand for three-dimensional space. For each positive number r, consider the sphere of radius r centered at the origin and let S_r denote the segment of the sphere in the first octant. Now if a mode of vibration in the cubic box with wavelength λ given by Eq. (51) is associated with a point (k, l, m) in R^3 of distance $r = \sqrt{k^2 + l^2 + m^2}$ from the origin, we have from Eq. (51) that

$$\lambda = 2L/r. \tag{52}$$

In terms of frequency $\nu = c/\lambda$ (c is the speed of light, the speed at which we assume electromagnetic waves propagate), we can rewrite Eq. (52) as

$$r = 2L\nu/c. \tag{53}$$

Let us now consider an interval of numbers $[r, r + \Delta r]$, and ask how many points (k, l, m) with integer coordinates lie in the region $Q(r, \Delta r)$ between S_r and $S_{r+\Delta r}$. If r is very large, which it is for the very short wavelengths in blackbody radiation, then this number of points can be approximated reasonably by the volume of $Q(r, \Delta r)$. To see this, imagine a large region of space filled with unit cubes centered at the points in the region with integer coordinates. These cubes fill space and each contains exactly one point with integer coordinates. The larger the region, the better we can approximate its volume with such unit cubes. In other words, points with integer coordinates occupy space with a density of one per unit volume.

Now the volume of $Q(r, \Delta r)$ can be approximated by the surface area of S_r times the thickness of $Q(r, \Delta r)$. So we can write that the number of integer points (k, l, m) in $Q(r, \Delta r)$ (one-eighth of a spherical shell) is

$$N_r = 4\pi r^2 \Delta r/8. \tag{54}$$

Each point with integer coordinates determines a mode of vibration of a particular wavelength, although it is clear from Eq. (51) that several points (modes) can correspond to the same wavelength. In addition, in electromagnetic radiation each mode can be polarized in one of two ways, so that we must double the right-hand side of Eq. (54) to account for all possible modes of vibration corresponding to $Q(r, \Delta r)$. Now let us use Eq. (53) to rewrite Eq. (54) in terms of frequency, replacing Δr with $(2L/c)\Delta\nu$, to obtain

$$N\nu = 8\pi\nu^2 L^3 \Delta\nu/c^3. \tag{55}$$

Finally, we allow $\Delta\nu$ to become infinitesimally small and divide by L^3 to obtain a density function giving the density of modes of vibration of the electromagnetic field per unit volume in the blackbody cavity as a function of frequency:

$$\phi(\nu) = 8\pi\nu^2/c^3. \tag{56}$$

The next step is to consider a unit volume of the cavity in thermal equilibrium at temperature T and the total electromagnetic energy $I(\nu, T)$ due to radiation of frequency ν. Then we partition this total energy into discrete packets of value $E(\nu, T)$ and distribute these packets among the modes associated with ν according to Eq. (56) to obtain a spectral density function for energy:

$$I(\nu, T) = \phi(\nu)E(\nu, T) = 8\pi\nu^2 E(\nu, T)/c^3. \tag{57}$$

Thus, we arrive once again at Eq. (22), this time without recourse to classical theory of linear oscillators. To be sure, this step is no less bold than Planck's original "act of desperation" cited in Section III.D. Debye was encouraged to use this assumption of quantized energy by the derivation of Eq. (34) in Einstein's theory of specific heat.

To obtain Planck's radiation law it is now necessary only to consider each mode of vibration to contribute the quantized value $E(\nu, T) = h\nu/[\exp(h\nu/kT - 1]$.

As we mentioned above, Debye's derivation of the radiation law is fundamentally more attractive than Planck's because, by avoiding arguments based on classical harmonic oscillators, it is possible to avoid the inconsistency in Planck's theory cited by Einstein. By the same token, this derivation is noteworthy for its lack of any mechnical model to account for the radiation. It is necessary in Debye's approach only to break up the variable, energy, into discrete quantum values and distribute those values over modes of vibration of a "vibrating" electromagnetic field. The quantum theory, the assumption that the physical quantity, energy, is a discontinuous variable, appears here as a fundamental law of nature, and the role of the classical mechanical model as nature's fundamental building block is thereby diminished.

Debye's work is noteworthy for another reason. It outlines the basic pattern that characterizes all applications of what we have called the quantum theory. That is, to apply the quantum theory to a physical phenomenon, first find a description of the phenomenon that involves both energy E and frequency ν. Second, set energy proportional to frequency, $E = h\nu$, where h is Planck's constant. All applications of early-20th-century quantum theory are variations on this theme. If $E = mc^2$ has become the internationally recognized equation symbolizing the theory of relativity, then $E = h\nu$ might well deserve the same status as the symbol of quantum theory.

IV. THE BOHR ATOM

A. The Planetary Atom

Between 1910 and 1913 an important development in physics was reported from Manchester, England. Ernest Rutherford had studied the results of bombarding a thin metallic foil with particles emitted by a radioactive substance. By observing how these particles were scattered after they hit the foil, Rutherford was able to propose a mechanical model of the atoms in the foil that could account for the scattering effect.

Rutherford's atom consists of a nucleus at the center of an electric field and a system of electrons that rotate about the nucleus in regular orbits, like planets around a sun. The electrons have small mass compared with the nucleus, and each one carries a negative electric charge e, today considered to have a value of about -1.6×10^{-19} coulomb (C). Since the atom is electrically neutral, the nucleus carries a positive charge $Z|e|$, where Z is the number of electrons in orbit. We call Z the atomic number of the atom.

This planetary model was considered quite a success, although it contained a fatal flaw, recognized even by Rutherford himself. The flaw is easy to see. A particle rotating in an orbit must have an acceleration toward the nucleus at every instant, or else it must fly off in a straight line. According to classical laws of accelerating charged particles, therefore, rotating electrons must constantly radiate energy in the form of electromagnetic radiation, as we mentioned in Section I.C. This loss of energy must result in a decrease in the orbital radius of the electron at such a rate that the electron must very quickly fall into the nucleus.

Thus, Rutherford's model is untenable, at least for atoms having only one electron, such as hydrogen. Atoms with many electrons might avoid collapse because of complicated interaction among the orbiting electrons, but no such escape clause can exist in Rutherford's theory for the hydrogen atom.

Niels Bohr, a Danish physicist who had visited Rutherford's laboratory in Manchester in 1912, applied the quantum theory to the planetary model of the atom. Bohr's theory can be conveniently outlined in four postulates. Like Planck's postulate about harmonic oscillators, some of Bohr's postulate were controversial because they were ad hoc, without explanation based on classical physics. Like Planck's postulate, however, they resulted in experimentally verifiable calculations, and so they were difficult to ignore. Furthermore, his postulates not only addressed the question of why atoms do not collapse but also introduced new mixtures of ideas relating waves to particles that were to become general principles of physics. Let us now examine Bohr's theory of the atom by paraphrasing his postulates.

B. Bohr's Postulates

1. Postulate I

Atoms exist in states of equilibrium, with electrons rotating in prescribed orbits about the nucleus, but, contrary to classical laws of electrodynamics, they do not radiate energy while in these orbits. The orbiting electrons are subject to classical laws of mechanics while in these states, however, so each possesses a mechanical energy (potential plus kinetic) E, determined by laws of orbiting bodies. These states of equilibrium, or nonradiation, are called "stationary states" of the atom.

2. Postulate II

Atoms can be made to change stationary states discontinuously in violation of classical laws of mechanics.

3. Postulate III

During transition from one stationary state to another, atoms emit or absorb quantities of energy (energy quanta) in "bundles" characterized by frequency. The frequency ν of a quantum of radiant energy emitted or absorbed in a change from one stationary state of mechanical energy E_1 to an adjacent state of energy E_2 is related to Planck's constant h by

$$\nu = (E_1 - E_2)/h. \tag{58}$$

Note the pattern: energy $= h\nu$, now becoming the hallmark of quantum theory.

Let us pause here to note one idea in this third postulate that foreshadows quantum mechanics. We can imagine

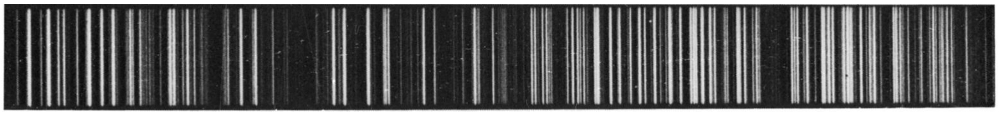


FIGURE 3 Emission spectrum of iron gas.

a bundle of radiant energy emitted from an atom, racing through space with many other bundles in a wavelike beam of radiation, and each bundle carries a message about the frequency of the beam according to Eq. (58). This idea was already well known for Einstein's light quanta, of course, but now it appears again in another context. The generalization of the idea of these "wave–particles" is one of the cornerstones of quantum mechanics.

Bohr's fourth postulate is directly related to the application of Rutherford's model to the study of atomic spectra: radiation emitted from an electrically charged gas. When an electric charge is passed through a gas, neon, for example, the gas emits electromagnetic radiation, predominantly red light in the case of neon. The radiation emitted covers a continuous set of frequencies, but it is much stronger at a certain discrete set of frequencies than it is at others, giving rise to the terms *continuous* and *discontinuous* spectra, the latter being the set of frequency values at which radiation is very strong. We shall again use the more traditional word *discontinuous* in place of *discrete*. If we pass the beam of radiation from a glowing gas through a prism that separates the waves according to frequency, we can observe part of the spectrum, the visible part, as in Fig. 3. The light lines indicate intense light at frequencies in the discontinuous part of the spectrum.

The spectral lines were the object of much research near the end of the 19th century. This research resulted in empirical formulas relating the frequencies of the discontinuous spectra to sets of positive integers. One such formula, developed by a Swiss schoolteacher, Johann Balmer, and reformulated by Janne Rydberg in 1890 (Rydberg claimed originality) can be written

$$\nu = R(1/n^2 - 1/m^2), \tag{59}$$

where R is a constant, known as Rydberg's constant, n and m are positive integers (with $m > n$), and ν is the frequency of a line in the discontinuous spectrum. For $n = 2$ and $m = 3, 4, 5, \ldots,$ frequencies given by Eq. (59) had been observed for the spectrum of hydrogen gas.

Let us now move to Bohr's fourth postulate and see how it leads to an explanation of empirical formula (59). Consider an electron in circular orbit of radius r in an atom of atomic number Z. At each point in the orbit the force on the electron due to the attraction of the nucleus is

$$F = Ze^2/r^2. \tag{60}$$

We then apply classical laws of mechanics [Eq. (8)] to arrive at the kinetic energy of the electron:

$$\text{KE} = Ze^2/2r. \tag{61}$$

The potential energy of the electron at every point in the orbit due to the attraction of the nucleus is

$$\text{PE} = -Ze^2/r. \tag{62}$$

This is the negative of the amount of work required to remove the electron from its orbit to a theoretical infinite distance from the nucleus. Thus, the total mechanical energy is

$$E = \text{KE} + \text{PE} = -Ze^2/2r. \tag{63}$$

The absolute value of E is the (positive) amount of energy required to bind an electron in its orbit, and we shall denote it, as is customary, by W. We then observe that the value of W equals the kinetic energy of the electron given by Eq. (61). Then, writing that kinetic energy in its classical form in terms of the rotational frequency ω of a particle of mass m moving in circular orbit of radius r, we have

$$W = |E| = \text{KE} = 2\pi^2 mr^2\omega^2. \tag{64}$$

This formula for binding energy was adjusted by Bohr to account for noncircular orbits and rotating nuclei, but our simplified form is sufficient to continue with our story of the Bohr atom.

The next step is based on the pattern characteristic of applications of quantum theory mentioned at the end of Section III. We seek a relation between energy W and orbital frequency ω and Planck's constant h. Bohr actually provided several plausibility arguments leading to the fourth postulate, some more convincing than others. We shall follow two of those arguments, one because it is simple, the other because it is based on an important philosophical principle. Remember, however, that these are just heuristic arguments. All of Bohr's postulates are ad hoc assumptions.

We begin the first heuristic argument by establishing a relation between orbital frequency ω and the frequency ν of radiation emitted during a change between stationary orbits. Assume that in the process of binding a free electron to the nucleus a binding energy quantum of frequency ν is required. If the orbital frequency resulting from the binding process is ω, then ν should be the average of ω for the orbiting electron and zero for the free electron. Thus,

$$\nu = \omega/2. \tag{65}$$

Now we can state Bohr's fourth postulate.

4. Postulate IV

An atom with one electron can exist in stationary states of equilibrium indexed by natural numbers $n = 1, 2, 3, \ldots$. In the state associated with number n, the electron orbits about the nucleus and the mean value of its kinetic energy in that orbit is given by

$$\mathrm{KE} = nh\omega/2, \tag{66}$$

where ω is its orbital frequency and h is Planck's constant. Note that this postulate is again a variation on the theme $E = h\nu$.

C. The Correspondence Principle

Next we shall examine a second argument leading to Eq. (66) based on what is now called the correspondence principle. We begin by assuming that the energy W for binding an electron into stationary orbit of frequency ω is proportional to $h\omega$. We then suppose that there is an orbit of lowest binding energy $W = \alpha h\omega$, where α is a proportionality constant, and that all other orbits have binding energies $W_1, W_2, \ldots$ given by

$$W_n = n\alpha h\omega, \tag{67}$$

where n is a natural number called the quantum number for the stationary state. The state with $n = 1$ is called the ground state of the atom.

Let us solve Eq. (64) for ω^2 to obtain

$$\omega^2 = W/2\pi^2 mr^2 \tag{68}$$

and replace r^2 using Eq. (61) and the fact that $W =$ KE to obtain

$$\omega^2 = 2W^3/\pi^2 me^4 Z^2. \tag{69}$$

Now if we index each orbit by its quantum number n to rewrite Eq. (69) using ω_n and W_n in place of ω and W, we can substitute Eq. (67) into Eq. (69) and solve for ω_n:

$$\omega_n = \pi^2 me^4 Z^2/2\alpha^3 n^3 h^3 \tag{70}$$

Using Eq. (67) once more, we obtain

$$W_n = n\alpha h\omega_n = \pi^2 me^4 Z^2/2\alpha^2 n^2 h^2. \tag{71}$$

Now let us use postulate III and consider the quantum of radiation with frequency ν emitted during a change of state from one with quantum number $n+1$ to the state with quantum number n. This is a change from a state of smaller binding energy W_{n+1} to one of higher binding energy W_n (with smaller orbit), which corresponds to a decrease in mechanical energy from a negative value E_{n+1} to a greater negative value E_n.

Then from Eq. (58) we have that

$$\begin{aligned}\nu &= (E_{n+1} - E_n)/h \\ &= (-W_{n+1} + W_n)/h,\end{aligned} \tag{72}$$

which from Eq. (71) gives us

$$\begin{aligned}\nu &= \frac{\pi^2 me^4 Z^2}{2\alpha^2 h^3}\left(\frac{1}{n^2} - \frac{1}{(n+1)^2}\right) \\ &= \frac{\pi^2 me^4 Z^2}{2\alpha^2 h^3}\left(\frac{2n+1}{n^2(n+1)^2}\right).\end{aligned} \tag{73}$$

This brings us to the correspondence principle:

> For very large quantum numbers n, the quantum theory frequency of radiation ν should correspond to the classical theory frequency ω of radiation from a charged particle in a circular orbit of orbital frequency ω.

In other words, the frequency ν given by Eq. (73) should equal the frequency ω_n given by Eq. (70) for large values of n.

Now for large values of $n, (2n+1)/n^2(n+1)^2$ is approximately equal to $2/n^3$. If we therefore replace $(2n+1)/n^2(n+1)^2$ in Eq. (73) with $2/n^3$ and use the correspondence principle to set $\nu = \omega_n$, we arrive at $\alpha = \frac{1}{2}$. We now take another bold step by declaring that what is true for large n must be true for all n. This will complete our second heuristic argument leading to postulate IV, for if the kinetic energy KE of a rotating electron in an orbit of quantum number n is to equal the binding energy W_n, we have from Eq. (67) (recall that $\alpha = \frac{1}{2}$) that

$$\mathrm{KE} = nh\omega/2. \tag{74}$$

The correspondence principle has been generalized and restated many times since it was first applied by Bohr in 1913. A form of this principle is another one of the cornerstones of quantum mechanics. That is why we have traced this last argument leading to Eq. (74).

D. Consequences of Bohr's Theory

First, note that we can rewrite Eq. (74) in terms of angular momentum l of an orbiting electron with kinetic energy KE. From Newtonian dynamics it can be shown that $l = \mathrm{KE}/\pi\omega$ for orbital frequency ω, and so from Eq. (74) we obtain

$$l = nh/2\pi \tag{75}$$

for stationary orbit with quantum number n. In other words, the quantization of the kinetic energy is equivalent to the quantization of the angular momentum of the orbiting electron. Though it would be much easier to derive Eq. (74) by simply assuming the quantization of angular momentum, that derivation completely avoids the correspondence principle and obscures a major philosophical point about Bohr's contribution to quantum theory.

Let us note also the close resemblance between Bohr's formula (73) and Rydberg's formula (59). As we mentioned, Bohr adjusted his value for mass *m* in Eq. (73) to account for the combined mass of nucleus and electron. Setting $\alpha = \frac{1}{2}$ in Eq. (73) gave Bohr an equation of the form of Eq. (59) from which he could calculate the value $R = 2\pi^2 me^4 Z^2/h^3$.

For hydrogen, where $Z = 1$, this formula provided excellent agreement with the experimentally determined value for *R*, using the best-known value for *h* at the time. Another major victory had been scored for quantum theory.

V. TRANSITION TO QUANTUM MECHANICS

There were several important victories for quantum theory between 1913 and 1925. None of them, however, provided new fundamental principles. So it is at this point that we conclude our discussion of quantum theory with a brief look at some of the steps that were required to make the transition from quantum theory to quantum mechanics.

Recall that what we have been calling quantum theory is really a collection of theories applied to different phenomena by mixing variable amounts of classical physics and quantum hypotheses. The three themes of quantum theory—the quantization of energy and the probabilistic behavior of energy quanta, the wave–particle nature of some matter, and Planck's constant—formed an interrelated set of ideas that lacked a universality and coherence necessary for them to constitute a scientific theory. Also lacking was a system of mathematical expressions common to all applications of quantum theory from which one could calculate the values of quantities observed in experiments.

Quantum mechanics, like Newtonian mechanics, was born of the necessity to bring mathematical clarity and order to the chaos of observations of the physical universe. Although Newtonian mechanics brought order to a set of observations of the continuous, predictable macroscopic world, it was inadequate to deal with the new chaos created by quantum theories of the discontinuous, unpredictable microscopic world.

One step in the transition from quantum theory to quantum mechanics was a philosophical one. Recall that in our discussion of Newton's contributions to science (Section I) we cited, as one of the most important, the notion that explanations of hidden, unobserved events come from precise measurements of observed events. In fact, it is this deductive power that some people equate with science itself. Yet Werner Heisenberg in 1925 saw that it was the universal application of this notion that stymied the development of a quantum mechanics. By first accepting the philosophical viewpoint that the only quantities of physics are the observable ones, he was able to produce his kinematics of quantum theory, a calculus of "observables." We now call it matrix mechanics, although Heisenberg never used the word *matrix* to describe the rectangular arrays of numbers that appeared in his equations as the observables.

One of the results of Heisenberg's mechanics was a calculation that involved Planck's constant in a profound way. With each observable Heisenberg defined a quantity called the uncertainty of the observable. He then showed that, for certain pairs of observables, the product of their uncertainties is at least as large as Planck's constant. Two observables related in this way are called *incompatible observables*. A consequence of this "uncertainty principle" is that the reduction of the uncertainty in one of the observables necessarily implies an increase in the uncertainty of the other. Heisenberg took these uncertainties to be a fundamental fact of nature, not a consequence of the inaccuracy of the measuring devices of physicists. Thus, he took one step farther Planck's reluctant acceptance of the fundamentally probabilistic behavior of oscillators and the implied uncertainty of that behavior. Heisenberg's complete break with the classical Newtonian physics of certainty sparked years of research that continues to this day, not only in physics, but also in philosophy, logic, and mathematics.

Another step in the transition to quantum mechanics was the theory of wave mechanics developed by Erwin Schrödinger. Working at the same time as Heisenberg in late 1925, but completely independently, Schrödinger produced a parameterized set of partial differential equations that had solutions only for a discrete set of values of the parameter. The equations involved a differential operator, and Schrödinger was able to show that his equations could be applied to any physical system by choosing appropriate differential operators. He then claimed that, after an operator was correctly chosen to represent a particular system, the discrete parameter values for which the resulting equation had solutions were all the possible values of energy for that system. Here, then, was a universally applicable mathematical formula that helped change quantum theory into quantum mechanics.

Although Schrödinger did not supply a convincing theoretical foundation for his equations, he was able to show in 1926 that his equations were mathematically equivalent to Heisenberg's matrix mechanics. There was no doubt in anyone's mind that the resulting combined theory, quantum mechanics, was a major scientific achievement. There was considerable doubt in the minds of some, however, Einstein most prominently, that the new mechanics, with its philosophical roots in the physics of uncertainty, was as universally applicable as its proponents claimed. Although controversy over its range of applicability persists to this day, quantum mechanics was the final step that brought Max Planck's "desperate act" to explain blackbody radiation to the status of a full-fledged scientific theory.

VI. AFTER QUANTUM MECHANICS: MODERN DEVELOPMENTS IN FOUNDATIONS OF QUANTUM PHYSICS

The quantum mechanics developed by Schrödinger and Heisenberg provided mathematical formulas with which to calculate physical values but lacked theoretical underpinnings. These were supplied by Max Born, Niels Bohr, and many others in the late 1920s and early 1930s. The theories, however, were controversial, and remain so to this day. While quantum mechanics has been successful in predicting the outcomes of some delicate experiments with photons, electrons, and other nuclear particles, some crucial predictions have not been experimentally confirmed. These predictions lie at the heart of the quantum theory controversy that we will explore in this section.

The most famous criticism of the theoretical foundations of quantum mechanics was published in 1935 by Albert Einstein, Boris Podolsky, and Nathan Rosen in a paper now called the "EPR paper" after the names of its authors. The paper describes a "thought experiment"—an experiment using an apparatus which could not be built at the time but which could test the theoretical predictions of quantum mechanics. The authors argued that the theory of quantum mechanics was incomplete because it contained no counterpart for an "element of reality" present in the objects in their apparatus. They were quite precise in their definition of an "element of reality":

> If, without in any way disturbing a system, we can predict with certainty . . . the value of a physical quantity, then there exists an element of physical reality corresponding to this physical quantity.

We shall describe a simplified version of the EPR thought experiment. Although our presentation is faithful to the ideas in the EPR paper, we have taken advantage of the 65 years scientists have had to study the paper to reframe the discussion for greater clarity.

Two quantum mechanical objects (photons or electrons, for example) are sent hurtling through space in opposite directions by a firing device. A certain measurement is performed on Object 1 (traveling to the left in the diagram) at collector A, and a similar measurement on Object 2 at collector B (see Fig. 4).

At each collector is a dial which can be set at any angle from 0 to 360 degrees. We shall explain the use for the dial later. The measurements are arranged so that for every setting of the dial, every time a pair of objects is fired, when an object reaches its collector one of two outcomes is recorded: "yes" or "no."

It is theoretically possible to design the EPR apparatus using special pairs of objects so that the outcomes of the measurements at the two collectors are correlated if the predictions of quantum mechanics are correct. Specifically, quantum mechanics predicts the following correlation:

> when the dials are set at the same angle, for each pair of objects fired, if a measurement on Object 1 at collector A registers "yes," then one can predict with certainty that the measurement on Object 2 at B will also register "yes," *whether or not the measurement at B is actually made.*

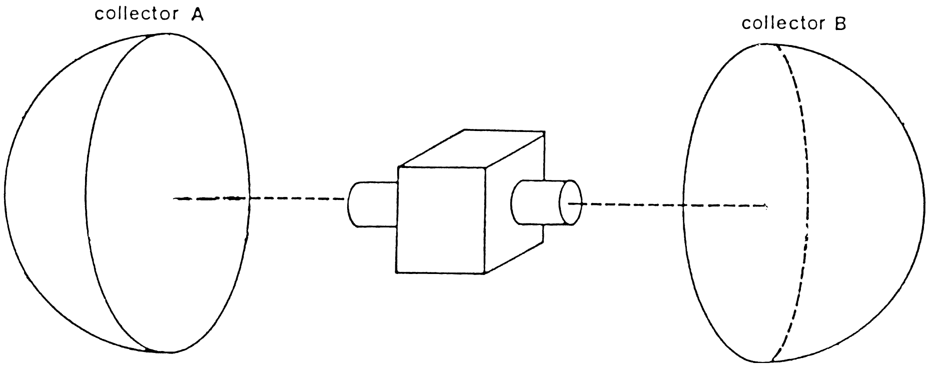


FIGURE 4 An EPR apparatus.

This implies that Object 2 has an "element of physical reality." It is carrying a code of some sort that determines how the measurement at B will turn out before the measurement takes place. The roles of Objects 1 and 2 could have been interchanged in our description, so, if we conclude that Object 2 is carrying a code, it is safe to assume that both objects carry a code as they race toward their respective collectors.

Einstein, Podolsky, and Rosen cited a basic tenet of quantum mechanics, which asserts that the objects in the EPR apparatus *cannot carry codes*. The theory of quantum mechanics requires that these objects behave probabilistically, and that *only a measurement* can reduce the probabilism to a state of certainty. The assumption that the objects cannot carry outcome-determining codes before they are measured is crucial to the theory of quantum mechanics; without it the entire theory collapses.

We can see why the EPR authors called the theory of quantum mechanics "incomplete." Because, after measuring Object 1, one can predict with certainty the value of the "yes–no" property of Object 2 without in any way disturbing it, this property has an element of physical reality which has no counterpart in the theory of quantum mechanics.

There was one response to the EPR paper which Einstein dismissed immediately, but which has lingered to this day as an intriguing possibility. Perhaps the objects in the EPR apparatus have no codes when they are fired, but by measuring Object 1 at A we somehow endow *both* objects with a code. This possibility is consistent with the theory of quantum mechanics, because it allows that the codes are produced by the act of a measurement. Einstein called this notion "spooky action at a distance." This is because the apparatus could theoretically be arranged so that Object 2 was far away from Object 1 at the time of measurement at A, requiring a "spooky action" to carry a message about the measurement at A *faster than the speed of light* to Object 2 in time to be measured at B, contradicting the theory of relativity.

Einstein, Podolsky, and Rosen ended their paper by asserting that while they had shown the theory of quantum mechanics to be an incomplete description of physical reality, they believed that a complete, realistic theory could be found eventually.

Hundreds of books and papers have been written about the EPR dilemma, and for years theoreticians searched for the complete, realistic theory Einstein and others believed to exist. In 1964, however, nine years after Einstein's death, a Scottish mathematician, J. S. Bell, poured cold water on the search. He showed that no "realistic, local" theory to explain the EPR experiment can exist that is consistent with the experimental outcomes predicted by quantum mechanics. A "realistic" theory asserts that the objects in the EPR apparatus carry codes which determine how the "yes–no" measurements at every angle setting of the dials will come out, whether or not the measurements are actually made. A "local" theory is one that prohibits a measurement at one collector from sending a signal across space at speeds greater than the speed of light to the other object.

We emphasize that the inconsistency demonstrated by Bell is not between realistic, local theories and the current *theory* of quantum mechanics, but rather between realistic theories and the *experimental results* predicted by the current theory of quantum mechanics. Thus, even if the current theory of quantum mechanics is discarded, if EPR-type experiments result in the correlations predicted by the current theory, then Bell's proof shows that *no* realistic, local theory can be used to explain those correlations.

Bell's result can be understood by considering the dials on the EPR apparatus. Suppose we assume that there is a realistic, local theory describing the behavior of the object pairs. By that we mean that for each pair fired an element of reality is accounted for by a code carried by the objects which determines for every setting of the dials which outcome ("yes" or "no") each collector will record when the object gets to it. Let us set both dials at 0 degrees. We know that quantum mechanics predicts that with these settings the outcomes at the two collectors will be perfectly correlated: each firing will result either in "yes" at both collectors or "no" at both.

Now suppose we move the A dial to a setting $+n$ degrees for some small number n, say 2 or 3, and leave the B dial at 0. Then if 100 pairs of objects are fired, there might be a loss of correlation. Let us denote by M_1 the number of mismatches ("yes" at A and "no" at B, or vice versa) in 100 firings. Next, let us leave the A dial at $+n$, and move the B dial to $-n$, (say 357 or 358 degrees), and fire another 100 pairs with exactly the same codes as the first 100 had. Bell's results establish that as long as we assume that the objects are carrying codes, and that what is recorded at one collector cannot affect what is recorded at the other, then when the dials are set at $-n$ and $+n$ during 100 firings, the number of mismatches M_2 in this new situation cannot exceed $2M_1$. That is $M_2 \leq 2M_1$. The theory of quantum mechanics predicts, however, that the number of mismatches in the second set of firings will be more than twice the number of mismatches in the first set. That is, $M_2 > 2M_1$. The predictions of quantum mechanics and realistic, local theories are then in irreconcilable conflict.

Much modern experimental research is devoted to turning the EPR-type thought experiments into real experiments to establish whether or not the predictions of quantum mechanics actually attain. One such series of experiments was performed by a team of scientists headed by Alain Aspect at Orsay, France, between 1982 and 1988. These experiments provide evidence that the results predicted by quantum mechanics do in fact attain in the physical world as we know it. The evidence, however, is not incontrovertible. The Aspect experiments used object detectors which are the best that can be built with modern technology, yet are only about 15% efficient. That is, only about 15% (or fewer) of the measurements at each collector can be guaranteed to be correct. To rule out the possibility that the results in the Aspect experiments were merely a coincidence produced by experimental error, we would have to build detectors that are at least 80% efficient. Most experimenters do not see the possibility of building such detectors in the foreseeable future.

We see then that 90 years after Max Planck put forth his "desperate" quantum hypothesis, the theory behind quantum behavior is still not fully formulated. Current formulations are fraught with ambiguities and counter-intuitive hypotheses. They are at odds with centuries of Western thought, including Platonism, classical physics, and most religions. It is understandable that many people are reluctant to throw out all of this tradition merely because one peculiar model for describing subatomic phenomena is enjoying high predictive value at the moment. It is little wonder that the theories of quantum physics are the object of such intense scrutiny and debate in the worlds of science, mathematics, and philosophy.

VII. AN ADVANTAGE OF UNCERTAINTY: QUANTUM COMPUTING

Even though the problems with the quantum theory exposed by the EPR paper and Bell's Theorem caused a great stir among philosophers and those fascinated by the foundations of science, they did not stop physicists from using quantum mechanics to make many important discoveries throughout the second half of the 20th century. As the century drew to a close, however, a troubling clash between quantum theory and high-speed computers suddenly turned into what promises to be an exciting marriage.

In the 1980s and 1990s high-speed computers became faster and faster, in part because their components became smaller and smaller. Computer "bits," the devices used to store the "zero-one" information at the heart of all computations are becoming so small that they're beginning to approach atomic scale. And on that scale the laws of quantum physics begin to apply.

The most troublesome quantum law facing computer designers is Heisenberg's Uncertainty Principle. To say that a computer bit is "in a state 0 or state 1" is a classical notion—one which requires a realistic, local theory of the physics governing the bit. As we saw in Sections V and VI, however, realistic, local theories are at odds with quantum mechanics. Instead, on the quantum level it is essential to accept the fact that a computer bit exists in a probabilistic state. So the best we might be able to say is that a bit is "in a zero state *with a given probability*." While this situation might seem to spell doom to the idea of building computers on the quantum level, it turns out instead to present a promising new direction for computing.

To see how uncertainty can be used to advantage let's look at a geometrical conceptualization of the Uncertainty Principle. Recall from Section V that Heisenberg's matrix mechanics identifies pairs of incompatible observables. Two observables are incompatible if at every instant the more certain we are about the value of one of them, the more *un*certain we must be about the value of the other. To make precise this notion of uncertainty of values let's consider a physical system (a spinning electron, for example) and an observable O_1 which can take on two values a and b (say, "spin-up" and "spin-down" in a certain direction). We can represent the two values with a perpendicular set of axes. We say that at every instant in time the system "exists in a state Ψ," which we represent as a vector of length one on our axis system (see Fig. 5).

Then we make the statement which lies at the heart of the probabilistic description of quantum mechanics. We say that *when the system is in state* Ψ, *then a measurement of observable* O_1 *will yield the value a with probability* $|\lambda_1|^2$, which is the square of the length of the projection of the state vector Ψ onto the a axis. At the same time *the measurement will yield value b with probability* $|\lambda_2|^2$, which is the square of the length of the projection of the state vector Ψ onto the b axis (see Fig. 6). (Note that $|\lambda_1|^2 + |\lambda_2|^2 = 1$, because Ψ has length one. So the probability that a measurement will yield "either a or b" is one.)

If Ψ lies along the positive a axis, we'll say that the system is in state $\bar{a}$, and if it's along the positive b axis,

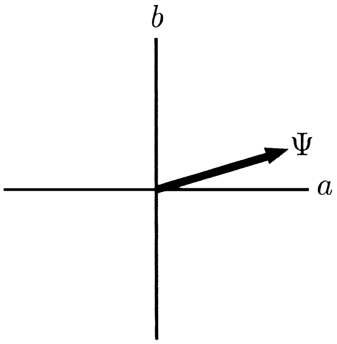


FIGURE 5 Two axes, representing possible values a and b for observable O_1, and a state vector Ψ.

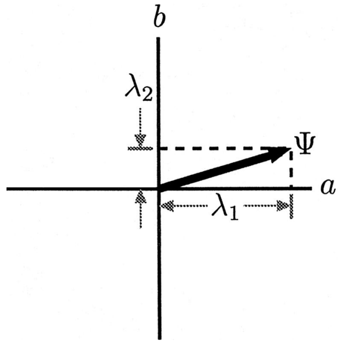


FIGURE 6 The squares of the lengths of the projections of the state vector Ψ give the respective probabilities that a measurement of observable O_1 will yield value *a* or *b* when the system is in state Ψ.

we'll say it's in state $\bar{b}$. If it's not along either axis, we'll say it's in a "*superposition*" of states $\bar{a}$ and $\bar{b}$.

Pay close attention to our description. When a system is in a superposition state, we don't say that it "has" the value *a* or the value *b*, and that we're uncertain as to which one it has. Rather, we say that the system has both values in some sort of mixture, with perhaps a greater probability of yielding one of these values rather than the other, *if we measure the observable while the system is in that state*. If we don't measure the observable, the system can continue to exist in the superposition state, and when we do measure it, the system jumps immediately into one of its two certain states, either $\bar{a}$ or $\bar{b}$. Because we don't say that the system "has" a value, our description is not a realistic, local theory, and we are sidestepping the EPR-Bell dilemma.

Now for the same physical system consider a second observable O_2, which can also take on two values, say x and y, and is incompatible with O_1 (for example, "spin-up" or "spin-down" in a different direction). Let's represent x and y with perpendicular axes rotated 45 degrees counter clockwise from those representing the values of O_1 (see Fig. 7).

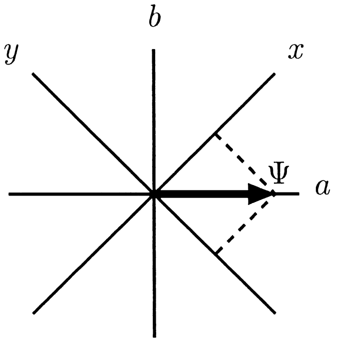


FIGURE 7 Two pairs of axes, representing possible values for two incompatible observables, and a state vector Ψ on the positive *a*-axis.

Now we can see a dramatic graphical representation of the Uncertainty Principle. Notice that if the state vector Ψ lies along the positive a axis, then its projections onto the x and y axes have length $\frac{1}{\sqrt{2}}$ (see Fig. 7). Hence if Ψ is a state for which O_1 has value a with certainty, then in that state a measurement of O_2 will result in x or y with probability of $\frac{1}{2}$ for each, a state of *maximal uncertainty!* Moreover, we can see the tradeoff between certainty for O_1 and certainty for O_2 as Ψ assumes various positions. That is, the more closely Ψ aligns itself with one of the lines—the x axis, for example—giving greater probability that a measurement of O_2 will yield x, the larger will be the projections of Ψ onto the a and b axes, increasing the uncertainty of O_1 for that state. There is no way to put the system in a state of zero uncertainty for O_1 and O_2 at the same time. That's what we mean when we say that O_1 and O_2 are incompatible observables.

This graphic representation of the Uncertainty Principle is a bit oversimplified. In practice physicists use complex numbers to describe superposition. But we won't need complex numbers to illustrate our example of a quantum computation.

How does a quantum computer take advantage of uncertainty? The answer lies in our assumption that a quantum physical system can have both values of a binary observable at the same time. We'll illustrate the power of quantum computing with a very simple example.

Let's consider a set of n objects $S=\{a_1, a_2, \ldots, a_n\}$, where n is an even number. Suppose we're given a function $f: S \to \{0, 1\}$. That is, f assigns either 0 or 1 to each a_j in S. We're told that f is either "constant" (i.e. $f(a_j) = f(a_k)$ for all j and k), or "balanced" (i.e. $f(a_j) = 0$ for exactly half the a_j's). Our problem is to determine whether f is constant or balanced. To see how quantum theory can help us with this problem, first let's look at it in the language of states and observables. To keep things simple we'll let $n = 2$.

At the heart of every classical computer is a binary bit, a physical device which can be in one of two states; magnitized or not, on or off, etc. Generally, the two states are labeled 0 and 1. Let's consider $\overline{a_1}$ and $\overline{a_2}$ as two states of a binary bit, which we'll depict as two perpendicular unit vectors

$$\overline{a_1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

and

$$\overline{a_2} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

in two-dimensional space. Then we'll represent our function f as a 2×2 matrix F with all off-diagonal entries

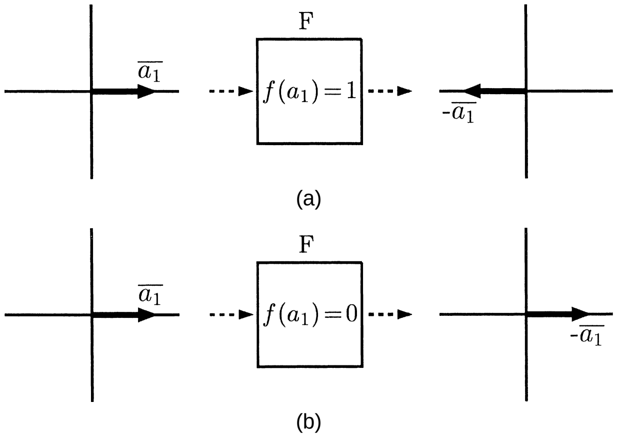


FIGURE 8 The function evaluator F flips state $\overline{a_1}$ only if $f(a_1)=1$.

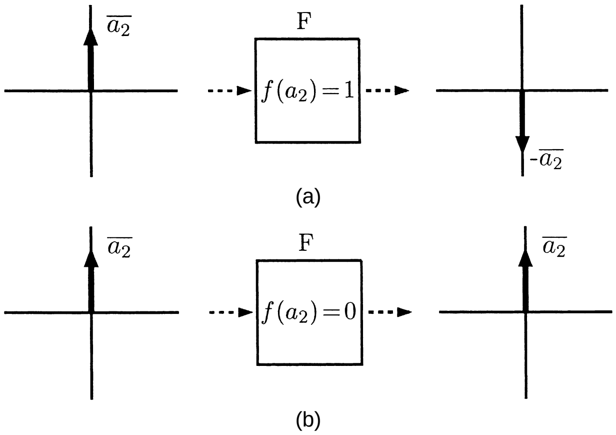


FIGURE 9 The function evaluator F flips state $\overline{a_2}$ only if $f(a_2)=1$.

zero, and diagonal entries $d_{jj}=(-1)^{f(a_j)}$ for $j=1, 2$. For example, if $f(a_1)=1$ and $f(a_2)=0$, then

$$F=\begin{bmatrix}-1 & 0\\ 0 & 1\end{bmatrix},$$

whereas if $f(a_1)=f(a_2)=1$, then

$$F=\begin{bmatrix}-1 & 0\\ 0 & -1\end{bmatrix}.$$

First, we'll see how a classical computer can determine whether a given f is constant or balanced. We can feed the state vector $\overline{a_1}$ into F to see if F flips the state vector. If $f(a_1)=1$, then

$$F\overline{a_1}=F\begin{bmatrix}1\\ 0\end{bmatrix}=\begin{bmatrix}-1\\ 0\end{bmatrix}=-\overline{a_1},$$

so $\bar{a}_1$ is flipped (see Fig. 8a). If $f(a_1)=0$, then

$$F\overline{a_1}=F\begin{bmatrix}1\\ 0\end{bmatrix}=\begin{bmatrix}1\\ 0\end{bmatrix}=\bar{a}_1.$$

(See Fig. 8b.)

Similarly, if $f(a_2)=1$, then the function evaluator F flips $\overline{a_2}$, otherwise it does not (see Fig. 9).

Clearly, to determine whether f is constant or balanced we need to feed both $\overline{a_1}$ and $\overline{a_2}$ to F. Feeding only one will not give enough information. We need to see if F flips both, neither, or only one of the two input states.

Turning now to a quantum computer, we see at its heart a "q-bit," a physical device which can be in one of two states (up or down, for example) or in a *superposition* of both states. Then we can prepare our input state in a superposition $\Psi=\frac{1}{\sqrt{2}}(\overline{a_1}+\overline{a_2})$, and feed that to F. What we get out is another superposition state Φ with its $\overline{a_1}$ and $\overline{a_2}$ components flipped or not, depending on the values of $f(a_1)$ and $f(a_2)$. For example, if f is constantly 0 ($f(a_1)=f(a_2)=0$), we get $F\Psi=\Phi=\Psi$. If f is constantly 1, then we get $F\Psi=\Phi=-\Psi$ (see Fig. 10).

On the other hand, if f is balanced, then $F\Psi$ is either Φ or $-\Phi$, as shown (see Fig. 11).

All that remains to do to determine whether f is constant or balanced is to measure the magnitude of the vector inner-product $P=|\Phi\cdot\Psi|$. If f is constant, then $P=1$, because $\Phi=\pm\Psi$. If f is balanced, then $P=0$, because Φ is orthogonal to Ψ. The inner product can be measured physically. The important thing to notice is that we needed to use the evaluator F only once, not twice. That is because we were able to feed it both states $\overline{a_1}$ and $\overline{a_2}$ simultaneously in a quantum superposition. And we should point out too that "applying the matrix F" involves only one physical operation in a quantum computer. It might mean firing a pulse through an array of polarized lenses. Do not confuse this operation with computing a matrix product classically. That takes many computer steps.

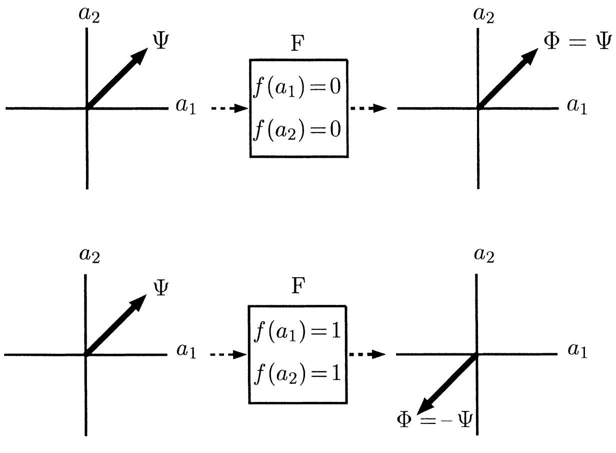


FIGURE 10 The superposition state Ψ goes to $\pm\Psi$ when f is constant.

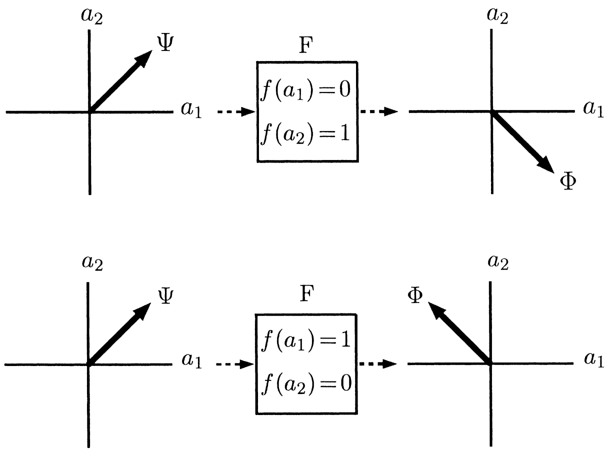


FIGURE 11 When fed to the evaluator F the superposition state Ψ goes to Φ when $f(a_1)=0$ and $f(a_2)=1$, and to $-\Phi$ when $f(a_1)=1$ and $f(a_2)=0$.

Using superposition to cut down evaluations of f from two to one might not seem like a revolutionary achievement. We can, however, apply the same principles to a set S with a very large (even) number n. Classically, we would need as many as $\frac{n}{2}+1$ executions of the evaluator F to determine whether f is constant or balanced in a worst-case scenario. Of course, we could be lucky when we start evaluating $f(a_1)$ and $f(a_2)$, and get $f(a_1)=1$ and $f(a_2)=0$, and know immediately that f is balanced. But if we evaluate $f(a_1)$, $f(a_2), \ldots, f(a_{\frac{n}{2}})$ and find that they're all equal, then still don't know whether f is constant or balanced. That's why we need $\frac{n}{2}+1$ evaluations in a worst case. Using a generalization of the superposition technique we used for the case $n=2$, however, we can design a quantum computer to determine whether f is constant or balanced with only one evaluation of a matrix F, even if $n=2^N$, where N is *any* natural number N. Cutting down the number of required evaluations of f from $\frac{n}{2}+1$ to 1 is a valuable achievement.

While we've illustrated how quantum theory might bring important advances to computing, we've left much unsaid. As we enter the 21st century, quantum computers exist mostly in theory. Although rudimentary quantum computers have been built, it isn't clear if it's within human reach to build one capable of handling serious problems. For one thing, it's very hard to create a superposition state in the physical world, and maintain it long enough to be of value before it breaks down, or is altered by spurious physical events. Further, physical operations, such as measuring the vector inner product, require instruments of very high reliability—much higher than we foresee in the near future. And finally, it is not yet known whether there exists a class of problems which are theoretically easier to solve with quantum algorithms than they are with classical algorithms. For example, the mathematician Peter Shor has invented a quantum algorithm which theoretically can be used to factor large numbers faster than all known classical algorithms. It has not been proved, however, that there does not exist a classical algorithm which can factor large numbers just as efficiently as the quantum algorithm. Factoring is a very important matter, because most computer security systems are based on the inability to factor large numbers quickly. Anyone who can build a high-powered factoring machine might be able to crack security systems around the world. That's one reason quantum computing is currently of great interest to physicists, mathematicians, and computer scientists.

SEE ALSO THE FOLLOWING ARTICLES

Atomic Physics • Celestial Mechanics • Chemical Thermodynamics • Electromagnetics • Mechanics, Classical • Photochemistry, Molecular • Quantum Mechanics • Radiation Physics • Statistical Mechanics

BIBLIOGRAPHY

Bell, J. S. (1964). "On the Einstein Podolsky Rosen paradox." *Physics* **1,** 195–200.

Cohen, D. W. (1989). "An Introduction to Hilbert Space and Quantum Logic," Springer-Verlag, New York.

Cropper, W. H. (1970). "The Quantum Physicists," Oxford Univ. Press, New York.

Einstein, A., Podolsky, B., and Rosen, N. (1935). "Can quantum mechanical description of physical reality be considered complete?" *Phys. Rev.* **47,** 777–780.

Hermann, A. (1971). "The Genesis of Quantum Theory," MIT Press, Cambridge, MA.

Jammer, M. (1966). "The Conceptual Development of Quantum Mechanics," McGraw-Hill, New York.

Kuhn, T. (1978). "Black-Body Theory and the Quantum Discontinuity," Clarendon Press/Oxford Univ. Press, New York.

Pais, A. (1982). "Subtle Is the Lord: The Science and Life of Albert Einstein," Oxford Univ. Press, London and New York.

Wheeler, J., and Zurek, W. (1983). "Quantum Theory and Measurement," Princeton University Press, Princeton, NJ. [This book contains reprints of the papers by Bell (1964) and Einstein, Podolsky, Rosen (1935).]

Relativity, General

James L. Anderson
Stevens Institute of Technology

GLOSSARY

Absolute object An element of a physical theory that affects but is itself unaffected by the other elements of the theory.

Binary pulsar Double star system, one of whose components is a neutron star.

Black hole A concentration of mass whose gravitational field is so strong that an event horizon forms around it.

Doppler shift Fractional change in frequency of light due to relative motion between source and observer.

Dynamical object An element of a physical theory whose behavior is affected by the other elements of the theory.

Electrodynamics Theory of electric and magnetic fields and of the interactions of the charged particles that produce them.

Event horizon Surface formed around a black hole through which nothing, including light, can penetrate from the inside. Any object that falls through such a surface is forever trapped inside it.

Free bodies Material objects that are not acted upon by external forces. Used in the construction of both Newtonian and special relativistic theories.

Galilean relativity The invariance of the laws describing physical systems with respect to Galilean transformations between observers moving with uniform velocity relative to each other.

Hubble constant Ratio of velocity of recession to distance of galaxies.

Light cone The surface formed by the light rays emanating from and converging on a point in space–time. Used in the construction of special relativistic theories.

Manifold A continuum of points characterized only by its global or topological structure.

Mapping Association of points of a space–time manifold with other points of this manifold.

Newtonian mechanics The basis for the description of physical systems obeying Newton's laws of motion.

Perihelion Point in the orbit of a planet when it is closest to the sun.

Planes of absolute simultaneity The collection of all point that are simultaneous with respect to one another. Used in the construction of Newtonian laws of motion.

Riemannian geometry Geometry in which the distance between neighboring points is defined by a metric and is quadratic in the coordinate differences between the points.

Space–time manifold Four dimensional manifold that underlies the construction of the various space–time descriptions of physical systems.

Space–time theories Physical theories that make use of the space–time manifold in the formulation.

Special relativity The invariance of the laws describing physical systems with respect to Lorentz transformations between observers moving with uniform velocity relative to each other.

THE GENERAL THEORY OF RELATIVITY is currently accepted as our best macroscopic description of the gravitation interaction that exists between all physical systems. It is also universal in that all physical systems are held to interact gravitationally. The predictions of general relativity differ both quantitatively and qualitatively from those of the Newtonian theory of gravity. Although the quantitative differences between the two theories is for the most part small, so that general relativity contains Newtonian gravity as an approximation, these differences have been extensively tested in the solar system and other astrophysical systems. The agreement between observation and theory is better than 0.5%. However, the qualitative predictions of the theory are its most exciting and challenging feature. Among others, the theory predicts the existence of gravitational radiation. Although this radiation has not been directly observed, the effects of its emission have been observed in the binary pulsar PSR $1913 + 16$ and agree with the predictions of the theory to within 3%. The theory also predicts the phenomenon of gravitational collapse leading to the creation of black holes. There is strong observational evidence that such objects exist in the universe. And finally, the general theory serves as the basis for our best description of the universe as a whole, the so-called hot big-bang cosmology.

I. SPACE–TIME THEORIES OF PHYSICS

A. The Space–Time Manifold

To understand the revolution wrought by the general theory it is useful to set it in a framework that encompasses it as well as the other two major structures of physics, Newtonian mechanics and special relativity. Basic to each of these structures is the notion of the space–time manifold consisting of a four-dimensional continuum of points. It is assumed only that any finite piece of this manifold can be mapped in a one-to-one manner onto a connected region of the four-dimensional Euclidean plane. Otherwise, these points are featureless and indistinguishable from each other, and the manifold as a whole is characterized only by its topological properties. While this manifold is not itself associated with any physical entity, it serves as the basis for the construction of the geometrical structures that are to be associated with such objects.

Since the points of the space–time manifold can be mapped onto the four-dimensional Euclidean plane, one can coordinatize the manifold by assigning to each point the coordinates of its image point in the Euclidean plane, x^μ, where the index μ takes on the values 0, 1, 2, 3. Because the points of the manifold are assumed to be indistinguishable, this mapping is to a large extent arbitrary and hence the coordinatization is also arbitrary. Depending on the topological structure of the manifold, it may be necessary to cover it with several overlapping coordinate "patches" to avoid singularities in the coordinates. If, for example, the manifold has the topology of the surface of a ball, it is necessary to employ two such patches to avoid the coordinate singularity one encounters at the pole when using the customary polar coordinates.

If the manifold is coordinatized in two different ways, for example, by using Cartesian or spherical coordinates, the coordinates used for one such coordinatization must be functions of those used for the other and vice versa. This relation is called a coordinate transformation. In order to preserve the continuity and differentiability of the manifold it must be continuous, nonsingular, and differentiable.

B. Geometrical Structures

In space–time theories, physical entities are associated with geometrical objects that are constructed on the space–time manifold. These objects can be of many different types. A curve can be associated with the trajectory of a particle and is specified by designating the points of the manifold through which it passes. This can be done by giving the coordinates of these points as functions $x^\mu(\lambda)$ of a monotonically varying parameter λ along the curve. Likewise, a two-dimensional surface could be specified by giving the coordinates of the surface as functions of two monotonic parameters, and similarly for three- and four-dimensional regions.

In addition to collections of points, one can introduce geometrical objects that consist of a set of numbers assigned to a point. These numbers are said to constitute the components of the geometrical object. The components of the velocity of a particle at a particular point in its trajectory would constitute such a collection. If the components are specified along a trajectory, a surface, or any other part of the space–time manifold, they are said to constitute a field. The temperature in a room can, for example, be associated with a one-component field. Likewise, the electromagnetic field surrounding a moving charge can be associated with a field consisting of six components.

The basic requirement that must be met in order that an object be geometrical is a consequence of the indistinguishability of the points of the space–time manifold. It is that under a coordinate transformation, the transformed components of the object must be functions solely of its original components and the coordinate transformation. This requirement is simply met in the case of curves, surfaces, and so forth; for example, given a curve, the transformed curve can be immediately calculated given the coordinate transformation.

An especially useful group of geometrical objects for associating with physical entities are those whose transformed components are linear, homogeneous functions of the original components. The simplest example is the single-component object called a scalar $\phi(x)$. Under a coordinate transformation, its transformed value is just equal to its original value. The other linear, homogeneous objects constitute the vectors, tensors, and pseudoscalars, pseudovectors, and pseudotensors. Vectors and pseudovectors are four-component objects (they come in two varieties called covectors and contravectors), while tensors and pseudotensors have larger numbers of components. There are also objects whose transformed components are linear but not homogeneous functions of the original components. Finally, there are objects whose transformed components are nonlinear functions of the original components, although such objects have not been used to any great extent. In all cases, however, the nature of the object is characterized by its transformation law.

It should be pointed out that not all objects one can construct are geometrical. The gradient of a scalar is a geometrical object while the gradients of vectors and tensors in general are not.

C. Laws of Motion

The basis for associating geometrical objects with physical entities is purely utilitarian—there is no general procedure for making this association. The numerical values that these objects can assume are taken to correspond to the observed values of the physical entities with which they are associated. Since not all such values can in general be observed, it is necessary to formulate a set of rules, called here laws of motion, that select from the totality of values a given set of geometrical objects can have, a subset that corresponds to possible observed values. Thus, if it is decided to associate a curve with the trajectory of a planet, one would have to discover a system of equations such as those obtained from Newton's laws of motion to select from the totality of all curves the subset that would correspond to actual planetary trajectories.

One requirement that one would like to be fulfilled by all laws of motion is that of completeness—every set of values allowed by them must, at least in principle, be observable. It is, after all, the purpose of the laws of motion to rule out unobservable sets of values. Nevertheless, there are problems associated with the imposition of such a requirement. There may, for example, be practical limitations on our ability to observe all the values allowed by a given set of laws. It is unlikely that we will ever be able to attain the energies needed to verify some of the predictions of the grand unified theories that are being considered today. However, if one of these theories correctly described all that we can observe about elementary particle interactions, we would not discard it because we could not directly observe its other predictions. More troubling, however, are limitations in principle on what we can observe. When applied to the universe as a whole, the general theory of relativity allows for many different possibilities, yet by its very nature we can observe only the universe in which we live. In the strict sense, then, the general theory should be considered incomplete. Nevertheless, it does correctly describe a vast range of phenomena, and so far there does not exist a more restrictive theory that does so. Therefore, probably, the best we can do is to require that a law not admit values that could be observed if they existed but do not.

D. Principle of General Covariance

However the laws of motion are formulated, they must be such as to be independent of a particular coordinatization of the space–time manifold. This requirement is called the principle of general covariance and was one of the basic principles employed by Einstein when he formulated the general theory. For the principle to hold, the laws of motion for a given set of geometrical objects must be such that all of the transforms of a set of values of these objects that satisfy the laws of motion must also satisfy these laws.

The principle of general covariance is not, as has sometimes been suggested, an empty principle that can be satisfied by any set of physical laws. If, for example, the geometrical object chosen to be associated with a given physical entity is a scalar field $\phi(x)$, then the only generally covariant law that can be formulated involving only this object is the trivial equation

$$\phi(x) = \text{const.} \tag{1}$$

In order to formulate a nontrivial law of motion for ϕ it is necessary to introduce other geometrical objects in addition to the scalar field. One possibility is to introduce a symmetric tensor field $g_{\mu\nu}(x)$ and its inverse $g^{\mu\nu}(x)$, which are related by the equation

$$g^{\mu\rho}g_{\rho\nu} = \delta^{\mu}_{\nu}, \tag{2}$$

where δ^{μ}_{ν} is the Kronecker delta, with values given by

$$\begin{aligned} \delta^{\mu}_{\nu} &= 1 \qquad \mu = \nu \\ &= 0 \qquad \mu \neq \nu \end{aligned} \tag{3}$$

and where the appearance of a double index such as ρ implies a summation over its range of values. One can then take as the law of motion for ϕ the equation

$$\left(\sqrt{-g}g^{\mu\nu}\phi_{,\mu}\right)_{,\nu} = 0, \tag{4}$$

where g is the determinant of $g_{\mu\nu}$ and $,\mu := \partial/\partial x^{\mu}$. The tensor field $g_{\mu\nu}$ cannot itself be given as a function of the coordinates directly since in that case Eq. (4) would not be generally covariant. Rather, it must in turn satisfy a law of motion that is itself generally covariant. If one requires that this law involve no higher than second derivatives of $g_{\mu\nu}$, it can be shown that there are in fact only three essentially different such laws for this object. One can form laws of motion for ϕ other than Eq. (4), but in each case it is necessary to introduce other geometrical objects for the purpose and to formulate laws of motion for them. In the general theory, $g_{\mu\nu}$ is associated with the gravitational field and hence couples to all other physical quantities through their equations of motion.

E. Absolute and Dynamical Objects

To understand the revolutionary nature of the general theory it is necessary to distinguish between two essentially different types of objects that appear in the various space–time theories. We call them absolute and dynamical, respectively. If the totality of values allowed by the laws of motion for some geometrical object, such as the tensor $g_{\mu\nu}$ introduced above are such that they can all be transformed into each other by coordinate transformations, we say that that object is an absolute object in the theory. This can occur if the law of motion for the object does not involve any of the other objects in the theory. The remaining objects in the theory are called dynamical objects. The electric fields associated with different charge distributions, for example, cannot in general be transformed into one another and hence must be associated with a dynamical object.

Given a theory with absolute objects, it is possible to coordinatize the space–time manifold so that they take on a specific set of values. In the case of the tensor $g_{\mu\nu}$, one of the three possible laws of motion mentioned above is such that every set of values allowed by it can be transformed so that, for every point of the space–time manifold, $g_{\mu\nu} = \text{diag}(1, -1, -1, -1)$. If these values are substituted into the other laws of motion they will no longer be generally covariant, but rather they will be covariant with respect to some subgroup of coordinate transformations. This subgroup will leave invariant the chosen values of the geometrical object (or objects) and will be called the invariance group of the theory. The structure of this group will be independent of which particular set of values allowed by the laws of motion is chosen for the absolute objects. If there are no absolute objects then the invariance group is just the group of all allowed coordinate transformations.

Absolute objects are seen to play a preferred role in a theory—their values are independent of the values of the dynamical objects of the theory while the converse is in general not the case. (If it is, the absolute objects become superfluous and can be ignored.) A theory with absolute objects thus violates a kind of general law of action and reaction. We will see that both Newtonian mechanics and special relativity contain absolute objects while the general theory does not.

II. NEWTONIAN MECHANICS

A. Absolute Time and Space

In his formulation of the laws of motion, Newton introduced a number of absolute objects, chief of which were his absolute space and absolute time. Absolute time corresponds to the foliation of the space–time manifold by a one-parameter family of nonintersecting three-dimensional hypersurfaces, which we call planes of absolute simultaneity. All of the points in a given plane are taken to be simultaneous with respect to each other. Furthermore, these planes are such that the curves associated with the trajectories of particles intersect each plane once and only once. The "time" at which such an intersection takes place is characterized by the value of the parameter associated with the plane being intersected. These planes are absolute in that their existence and structure are assumed to be independent of the existence or behavior of any other physical system in the space–time.

In Newtonian mechanics the interaction of particles is assumed to be instantaneous as in Newton's action-at-a-distance theory of gravity. Consequently, such interactions take place between the points on the trajectories that lie in the same plane of absolute simultaneity. As a consequence, these planes can be observed by giving an impulse to one member of a group of charged particles and noting where, on the trajectories of the other particles, the transmitted impulse acts.

Newton's absolute space corresponds to a unique three-parameter congruence of nonintersecting curves that fill the space–time manifold; that is, through each point of the manifold there passes one and only one such curve. Furthermore, each curve passes through one and only one point of each plane of absolute simultaneity. The existence of such a congruence would therefore imply that there exists a unique one-to-one relation between the points in any two planes of absolute simultaneity. The "location" of a space–time point would be characterized by the parameters associated with the curve of the congruence passing through it.

The notion of absolute space brings with it the notion of absolute rest: a particle is absolutely at rest if its trajectory can be associated with one of the curves of the congruence. However, unlike the planes of absolute simultaneity that are needed in the formulation of the laws of motion of material particles, these laws do not require the existence of the space–time congruence of curves that constitute Newton's absolute space, nor do they afford any way of detecting a state of absolute rest. This property of the Newtonian laws of motion is known as the principle of Galilean relativity. Furthermore, since the congruence is not needed in the formulation of these laws, we can dispense with it and hence with Newton's absolute space altogether as an unobservable element of the theory.

B. Free Bodies

In his setting down of the three laws of motion, Newton was careful to give the first law, "Every body continues in its state of rest, or of uniform motion in a right line, unless it is compelled to change that state by forces impressed upon it," as separate and distinct from the second law. He clearly did not consider it, as it is sometimes taken to be, a special case of the second law. In effect, the first law supposes a class of curves, the straight (right) lines, to exist in the space–time manifold. Furthermore, these curves correspond to the trajectories of a class of objects on which no forces act, namely, free bodies. As a consequence, these curves are absolute objects of the theory. Furthermore, they, like the planes of absolute simultaneity, are needed to formulate the laws of motion for bodies on which forces act.

C. Galilean Invariance

One can always coordinatize the space–time manifold in such a way that the parameter t, which labels the different planes of absolute simultaneity, is taken to be one of these coordinates. When this is done, the equation that defines these planes is simply

$$t = \text{const.} \tag{5}$$

Furthermore, the remaining coordinates can be chosen so that the equations of the curves associated with the trajectories of the free bodies are linear in t; that is, they are of the form

$$x_i = v_i t + x_{0i}, \tag{6}$$

where the index i takes on the values 1, 2, 3, and v_i and x_{0i} are constants. The constants v_i are the components of the "velocity" of the free body whose trajectory is associated with this curve and the x_{0i} are its initial positions. When expressed in terms of these coordinates, the laws of motion of Newtonian mechanics take on their usual form.

Since the planes of absolute simultaneity and the straight lines constitute the absolute objects of Newtonian mechanics and enter into the formulation of the laws of motion of all Newtonian systems, the subgroup of coordinate transformations that leave them invariant as a whole constitutes the invariance group of Newtonian mechanics. In addition to the group of spatial rotations and translations and time translations, this group consists of the Galilean transformations given by

$$x'_i = x_i + V_i t \tag{7a}$$

and

$$t' = t, \tag{7b}$$

where the V_i are the components of the velocity that characterize a particular transformation of the group. The requirement of invariance under this group of transformations is called the principle of Galilean relativity.

In terms of the primed coordinates, we see that the equations of a straight line (6) take the form

$$x'_i = v'_i t' + x_{0i}, \tag{8}$$

where the transformed velocity components v'_i are given by the Galilean law of addition for velocities:

$$v'_i = v_i + V_i. \tag{9}$$

III. SPECIAL RELATIVITY

A. Light Cones

The transition from Newtonian mechanics to special relativity in the early part of this century involved the abandonment of the Newtonian planes of absolute simultaneity and their replacement by a new set of absolute objects, the light cones. With the completion of the laws of electrodynamics by Maxwell in the middle of the last century it became evident that electromagnetic interactions between charged particles were not instantaneous but rather were transmitted with a finite velocity, the speed of light. This fact, coupled with the Galilean velocity addition law, made it appear possible that some electromechanical experiment could be devised for the detection of a state of absolute rest and thus reinstate Newton's absolute space. However, all attempts to do so, such as those of Michelson and Morley and Trouton and Noble, proved fruitless. In one way or another, these experiments sought to measure the absolute velocity of the earth with respect to this absolute space. Even though they were sensitive enough to detect a velocity as small as 30 km/sec, which is much less than the known velocity of the earth with respect to the galaxy, no such motion was ever detected.

Einstein realized that if all interactions were transmitted with a finite velocity there was no way objectively to observe the Newtonian planes of absolute simultaneity and that they, like Newton's absolute space, should be eliminated from the theory. It was his analysis of the meaning of absolute simultaneity and its rejection by him that distinguished his approach to special relativity from those of Lorentz and Poincaré. Since, however, unlike absolute space, the planes of absolute simultaneity were needed in the formulation of the laws of motion for material bodies, it was necessary to replace them by some other structure. The key to doing this lay in Einstein's postulate that the velocity of light is independent of the motion of the source. If this is the case, and to date all experimental evidence supports this postulate, the totality of all light ray trajectories form an invariant structure and can be associated with a corresponding family of three-dimensional surfaces in the space–time manifold, the light cones. Just as in Newtonian mechanics, where through each point there passes a unique plane of absolute simultaneity, in special relativity through each point there passes a light cone that consists of all of the points on the curves passing through this point that correspond to the trajectories of light rays.

In Newtonian mechanics the interaction of particles was assumed to take place between the points on the curves associated with their trajectories that lay in the same plane of absolute simultaneity. In special relativity this interaction is assumed to take place, depending on the type of interaction that exists between the particles, either between points that lie in the same light cone or between one such point and points in the interior of the light cone associated with this point. The electromagnetic interaction, for example, takes place between points lying in the same light cone. Consequently, these light cones can be observed by giving an impulse to one member of a group of charged particles and noting where, on the trajectories of the other particles, the transmitted impulse acts.

B. Free Bodies

In addition to the absolute light cones, special relativity assumes, like Newtonian mechanics, a family of curves, the straight lines, that are associated with the trajectories of free bodies. There is, however, an important difference between the two theories. In Newtonian mechanics any straight line that intersects all of the planes of absolute simultaneity is assumed to correspond to the trajectory of a free body. In special relativity, on the other hand, only the straight lines that correspond to free bodies with velocities less than or equal to the speed of light are assumed to correspond to observable free bodies. These straight lines are such that, given a point lying on one of them, the other points lying on it are either interior to or lie on the light cone associated with that point.

C. Lorentz Invariance

Together, the light cones and straight lines constitute the absolute objects of special relativity. It can be shown that one can coordinatize the space–time manifold in such a way that the points with coordinates x^μ lying on a straight line are given by the equations

$$x^\mu = v^\mu \lambda + x_0^\mu, \tag{10}$$

where the v^μ and x_0^μ are constants and λ is a monotone increasing parameter along the line. For the straight lines that correspond to the trajectories of free bodies, the v^μ are constrained by the condition that

$$\eta_{\mu\nu} v^\mu v^\nu \geqq 0 \tag{11}$$

where $\eta_{\mu\nu} = \text{diag}(1, -1, -1, -1)$. Provided that $\eta_{\mu\nu} v^\mu v^\nu > 0$, it is always possible to choose the parameter λ such that $\eta_{\mu\nu} v^\mu v^\nu = 1$. In this case the v^μ are said to constitute the components of the four-velocity of the particle and $\tau = \lambda$ is called the proper time along the line.

In addition to the form (10) for the straight lines, coordinates can be chosen so that the points x^μ lying on the light cone associated with the point x_0^μ satisfy the equation

$$\eta_{\mu\nu}\left(x^\mu - x_0^\mu\right)\left(x^\nu - x_0^\nu\right) = 0. \tag{12}$$

For points interior to this light cone the quantity on the left side of this equation is greater than zero, while for points exterior to it it is less than zero. In what follows, coordinates in which the straight lines and light cones are described by Eqs. (10) and (12) will be called inertial coordinates.

Since the light cones and straight lines are absolute objects in special relativity, the coordinate transformations that leave these structures invariant constitute the invariance group of special relativity. In an inertial coordinate system, these transformations have the form

$$x'^\mu = a_\nu^\mu x^\nu + b^\mu, \tag{13}$$

where α^μ_ν and b^ν are constants. The b^μ are arbitrary while the α^μ_ν are constrained to satisfy the conditions

$$\eta_{\mu\nu}\alpha^\mu_\rho\alpha^\nu_\sigma = \eta_{\rho\sigma}. \tag{14}$$

These transformations form a group, the inhomogeneous Lorentz group, each member of which is characterized by the 10 arbitrary values one can assign to the α^μ_ν and b^μ. This group contains, as subgroups, the three-dimensional rotation group and the group of spatial and temporal translations. It also includes the group of Lorentz transformations, now called Lorentz boosts. A boost along the x axis takes the form

$$\begin{aligned} x'^0 &= \gamma(x^0 + \beta x^1)\\ x'^1 &= \gamma(\beta x^0 + x^1)\\ x'^2 &= x^2\\ x'^3 &= x^3, \end{aligned} \tag{15}$$

where $\gamma = (1 - \beta^2)^{-1/2}$ and β is a parameter that characterizes the boost. These transformation equations take their more familiar form if we set $x^\mu = (ct, x, y, z)$ and similarly for x'^μ and $\beta = v/c$, where c is the velocity of light, in which case v is the velocity associated with the transformation. In special relativity, the Lorentz boosts replace the Galilean transformations of Newtonian mechanics just as the light cones replace the planes of absolute simultaneity and the requirement of invariance under this group is called the principle of Special relativity. Also, the Galilean law of addition for velocities, Eq. (9), is no longer valid. For a boost in the x direction, the transformed components v'_i of the velocity of a body are related to its original components v_i by the equations

$$v'_i = \delta(v_1 + v) \qquad v'_2 = \gamma^{-1}\delta v_2 \qquad v'_3 = \gamma^{-1}\delta v_3, \tag{16}$$

where $\delta = (1 + v_1 v/c^2)^{-1}$.

D. The Space–Time Metric

Equations (10) and (12) for straight lines and light cones are given in a special coordinate system in which they assume these simple forms. It is possible to write generally covariant equations for these objects by introducing a symmetric second rank tensor $g_{\mu\nu}$ of signature -2 together with its inverse $g^{\mu\nu}$. With its help, the light cones can be characterized by the surfaces $\phi(x) = 0$, where ϕ satisfies the covariant equation

$$g^{\mu\nu}\phi_{,\mu}\phi_{,\nu} = 0. \tag{17}$$

To construct an equation for a straight line we first introduce the Christoffel symbols $\left\{ {\mu \atop \rho\sigma} \right\}$ defined by

$$\left\{ {\mu \atop \rho\sigma} \right\} = \tfrac{1}{2} g^{\mu\nu}(g_{\rho\nu,\sigma} + g_{\nu\rho,\sigma} - g_{\rho\sigma,\nu}). \tag{18}$$

These quantities constitute the components of a geometrical object that is linear but not homogeneous. With their help, the equations for the coordinates $x^\mu(\lambda)$ of the points lying on a straight line can now be written as

$$d^2x^\mu/d\lambda^2 + \left\{ {\mu \atop \rho\sigma} \right\}(dx^\rho/d\lambda)(dx^\sigma/d\lambda) = 0, \tag{19}$$

where again λ is a monotone increasing parameter along the curve. One can choose λ so that $g_{\mu\nu}\, dx^\mu/d\lambda\, dx^\nu/d\lambda = 1$, in which case it is the proper time along the line. One can also use Eq. (19) to characterize the trajectories of light rays if one adds the condition that $g_{\mu\nu}\, dx^\mu/d\lambda\; dx^\nu/d\lambda = 0$. Such rays have the property that they serve as the generators of the light cones. Equation (19) is usually referred to as the geodesic equation since it has the same form as the equation for a geodesic curve, that is, a curve of minimum length connecting two points in a Riemannian space with a metric $g_{\mu\nu}$.

Having introduced the tensor $g_{\mu\nu}$, it now becomes necessary to construct a law of motion for it. In special relativity this law is taken to be

$$R_{\mu\nu\rho\sigma} = 0, \tag{20}$$

where the tensor $R_{\mu\nu\rho\sigma}$, called the Riemann–Christoffel tensor, is constructed from the tensor $g_{\mu\nu}$ according to

$$\begin{aligned} R_{\mu\nu\rho\sigma} = {} & \tfrac{1}{2}(g_{\mu\rho,\nu\sigma} + g_{\nu\sigma,\mu\rho} - g_{\mu\sigma,\nu\rho} - g_{\nu\rho,\mu\sigma})\\ & + g_{\alpha\beta}\left(\left\{ {\alpha \atop \mu\rho} \right\}\left\{ {\beta \atop \nu\sigma} \right\} - \left\{ {\alpha \atop \mu\sigma} \right\}\left\{ {\beta \atop \nu\rho} \right\}\right). \end{aligned} \tag{21}$$

It appears in the equation of geodesic deviation that governs the separation between two neighboring, freely falling bodies. When all of the components of $R_{\mu\nu\rho\sigma}$ vanish, this separation remains constant.

It can be shown that every solution of Eq. (20) can be transformed so that $g_{\mu\nu} = \eta_{\mu\nu}$ everywhere on the space–time manifold, in which case $g_{\mu\nu}$ is said to take on its Minkowski values. In a coordinate system in which $g_{\mu\nu}$ has this form, Eq. (17) becomes

$$\eta^{\mu\nu}\phi_{,\mu}\phi_{,\nu} = 0 \tag{17a}$$

It is seen that the surface defined by Eq. (12) satisfies this equation. Likewise, in this coordinate system, Eq. (19) reduces to

$$d^2x^\mu/d\lambda^2 = 0 \tag{19a}$$

and has, as its solution, the curves defined by Eq. (16).

Since $g_{\mu\nu}$ can always be transformed to its Minkowski values, it is seen to be an absolute object in the theory and the group of transformations that leave it invariant is again the inhomogeneous Lorentz group. In all respects $g_{\mu\nu}$ is equivalent to the straight lines and light cones of the theory. Furthermore, one can either construct laws of motion that employ inertial coordinates and that are covariant with respect to the inhomogeneous Lorentz group or construct generally covariant laws with the help of the $g_{\mu\nu}$. When $g_{\mu\nu}$ is transformed to take on the values $\eta_{\mu\nu}$, the latter equations reduce to the former.

The Riemann–Christoffel tensor arose in the study of the geometry of manifolds with Riemannian metrics. In such a geometry one defines a distance ds between neighboring points of the manifold with coordinates x^μ and $x^\mu + dx^\mu$ to be

$$ds^2 = g_{\mu\nu}(x)\, dx^\mu\, dx^\nu, \qquad (22)$$

where $g_{\mu\nu}$, a symmetric tensor field, is the metric of the manifold. The vanishing of the Riemann–Christoffel tensor can be shown to be the necessary and sufficient condition for the geometry to be flat; that is, the metric can always be transformed to a constant tensor everywhere on the manifold. Since the tensor $g_{\mu\nu}$ introduced above is an absolute object it is sometimes referred to as the metric of the flat space–time of special relativity. Although we will not need to make use of this geometrical interpretation, it will sometimes prove convenient to give an expression for ds as a way of specifying the components of $g_{\mu\nu}$.

IV. GENERAL RELATIVITY

A. The Principle of Equivalence

After Einstein formulated the special theory of relativity he turned his attention to, among other things, the problem of constructing a Lorentz-invariant theory of gravity. Newton thought of gravity as an action-at-a-distance force between massive bodies and as transmitted instantaneously between them. Since special relativity required a finite speed of transmission, Einstein sought to construct a relativistic field theory of gravity. The simplest object to associate with the gravitational field was a scalar field. However, a difficulty presented itself when he came to construct a source for this field. In the Newtonian theory, the gravitational attraction between bodies was proportional to their masses. In special relativity, however, energy has associated with it an equivalent mass through the relation $E = mc^2$. Consequently, Einstein argued, mass density by itself could not be the sole source of the gravitational field. At the same time, energy density could not be used since it is not, by itself, associated with a geometrical object in special relativity but rather with one component of a tensor field.

While thinking about the problem of gravity, Einstein was struck by a peculiarity of the gravitational interaction between bodies, namely, the constancy of the ratio of the inertial to the gravitational mass of all material bodies. In Newtonian mechanics, mass enters in two essentially different ways—as inertial mass in the second law of motion and as gravitational mass in the law of gravitational interaction. Logically, these two masses have nothing to do with one another. Inertial mass measures the resistance of a body to forces imposed on it while gravitational mass determines, in the same way as electric charge determines the strength of the electrical force between charged bodies, the strength of the gravitational force between massive bodies. Galileo was the first to demonstrate this constancy by observing that the acceleration experienced by objects in the earth's gravitational field was independent of their mass. In 1891, Eötvös demonstrated it to an accuracy of one part in 10^8. More recent determinations by groups in Princeton and Moscow have established this constancy to better than one part in 10^{11}.

While there was no explanation for this constancy, Einstein realized that it called into question the existence of one of the absolute objects of both Newtonian mechanics and special relativity, the free bodies. If indeed this ratio was a universal constant, then there could be no such thing as a gravitationally uncharged body since zero gravitational mass would then imply zero inertial mass. Einstein also realized that this constancy meant that it would be impossible to distinguish locally, that is, in a sufficiently small region of space–time, between inertial and gravitational effects through their action on material bodies. An observer in an elevator being accelerated upward with an acceleration equal to that produced by the earth's gravity would see objects fall to the floor of the elevator in exactly the same way that they fall on earth, that is, with an acceleration that is independent of their mass.

After this realization, Einstein made a characteristic leap of imagination. He postulated that it is impossible to distinguish locally between inertial and gravitational effects by any means. One of the consequences of this postulate, called by him the principle of equivalence, is that light should be bent in a gravitational field just as it would appear to be to an observer in an accelerating elevator. But if this is the case, the light cones of special relativity would no longer be absolute objects either, and this in turn would mean that the metric of special relativistic space–time would not be an absolute object.

The principle of equivalence however, implied even more. If inertial and gravitational effects are indistinguishable from each other locally, then one and the same object could be used to characterize both effects. Since it is the metric $g_{\mu\nu}$ that is responsible for the inertial effects one observes in special relativity, $g_{\mu\nu}$ should also be associated with the gravitational field. In effect, geometry and gravity became simply different aspects of the same thing. Actually, one never needs to interpret $g_{\mu\nu}$ as a metric. One can identify it solely with the gravitational field.

This identification has, as a consequence, that $g_{\mu\nu}$ must be a dynamical object since the gravitational field clearly must be such. Having recognized this fact, Einstein then turned his attention to the problem of constructing a law of motion for this object.

B. The Principle of General Invariance

In his attempts to construct a law of motion for $g_{\mu\nu}$, Einstein proposed that these laws should be generally covariant. However, we have already seen that the laws of motion of special relativity could be cast in generally covariant form with the introduction of a metric satisfying Eq. (20). But such a metric was absolute and Einstein wanted a law of motion for a dynamical $g_{\mu\nu}$. Consequently, what Einstein was really requiring was not general covariance but rather general invariance, that is, that the invariance group of the laws of motion should be the same as their covariance group, namely, the group of all arbitrary coordinate transformations. As we have argued above, this can be the case only if there are no absolute objects in the theory. The absence of absolute objects in the theory satisfies a version of Mach's principle which states that there should be no absolute objects in any physical theory.

Although Einstein did not use precisely the reasoning outlined above, it was his recognition of the preferred role played by the inhomogeneous Lorentz group in special relativity that was crucial to the development of the general theory of relativity. And although he formulated his argument in terms of the relativity of motion, it is clear that he was referring to the invariance properties of the laws of motion. His argument that all motion should be relative—hence the term general relativity—was really a requirement, in modern terms, that these laws should be generally invariant. This is not an empty requirement, as some authors have suggested, but rather severely limits the possible laws of motion one can formulate for $g_{\mu\nu}$.

C. Dynamic Laws

1. Field Equations

The search for a generally invariant law of motion for $g_{\mu\nu}$ occupied a considerable portion of Einstein's time prior to the year 1915. At one point he even argued that such a law could not exist. However, he did succeed in that year in finally formulating this law. If one requires that this law contain no higher than second derivatives of the $g_{\mu\nu}$ and furthermore that it be derivable from a variational principle, then there is, infact, essentially only one law that fills these requirements. This is in marked contrast to the situation in electrodynamics, where there are an infinite number of laws of motion for the vector potential A_μ that satisfy these requirements.

To formulate the law of motion for $g_{\mu\nu}$, we first construct from the Riemann–Christoffel tensor (21) the Ricci tensor $R_{\mu\nu}$, where

$$R_{\mu\nu} = g^{\rho\sigma} R_{\sigma\mu\rho\nu} \tag{23}$$

and the curvature scalar R, where

$$R = g^{\mu\nu} R_{\mu\nu}. \tag{24}$$

In terms of these quantities this law can be written as

$$R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R + \Lambda g_{\mu\nu} = \kappa T_{\mu\nu}, \tag{25}$$

where Λ and κ are constants and $T_{\mu\nu}$ is the energy–momentum tensor associated with the sources of the gravitational field.

In general, the components of the Riemann–Christoffel tensor will not vanish even when all of the components of the Ricci tensor do. As a consequence, it follows from the equation of geodesic deviation that the separation between neighboring freely falling masses will change with time. Since such changes appear due to tidal forces in manypracticle systems, the Riemann–Christoffel tensor is thus a measure of such forces and vice versa.

The so-called cosmological term $\Lambda g_{\mu\nu}$ was originally not present in the Einstein field equations. It was later added by him to obtain a static cosmological model with matter. When it was discovered that the universe was expanding and that there were solutions of the field equations without the cosmological term that fit the then available data, Einstein dropped this term and for the most part it has not been included in the field equations. However, recent data suggest that the expansion of the universe is accelerating instead of slowing down as it does in the older cosmological models. One way to take into account such a discovery would be to reintroduce the cosmological term in the field equations.

In addition to the law of motion for $g_{\mu\nu}$ it is necessary to formulate generally invariant laws of motion for the other geometrical objects that are to be associated with the physical quantities being observed. One way to do this is simply to take over the generally covariant form of the laws formulated for these objects in special relativity. The law of motion for the electromagnetic field, when this field is associated with a vector A_μ, can, for example, be written in the form

$$\left(\sqrt{-g}\, g^{\mu\rho} g^{\nu\sigma} F_{\rho\sigma}\right)_{,\mu} = 4\pi j^{\nu}, \tag{26}$$

where g is the determinant of $g_{\mu\nu}$, j^μ the current density associated with the sources of the electromagnetic field, and

$$F_{\mu\nu} = A_{\nu,\mu} - A_{\mu,\nu}. \tag{27}$$

Such laws of motion are said to involve minimal coupling to the gravitational field. It is also possible to construct laws of motion that do not couple minimally to the gravitational field. In the case of the electromagnetic field, for example, one could include a factor of $1 + R$, where R is the curvature scalar, inside the parentheses in Eq. (26). The only requirement these laws of motion should satisfy is that they reduce to their special relativistic form when the Riemann–Christoffel tensor vanishes.

2. Particle Dynamics

When Einstein proposed the general theory in 1915 he took the geodesic Eq. (19) to be the equation of motion for a test particle moving in an external gravitational field. Such a particle was assumed to have no effect on the sources of this external field. Since, however, even a test particle has a finite mass it cannot avoid having some small but finite effect on these sources. Hence, the geodesic equation yields at best an approximation to the true motion of both the test particle. Furthermore, this equation cannot be applied at all in the important case of the mutual interaction of bodies of comparable mass. And finally it cannot deal with gravitational radiation effects.

In 1939 Einstein together with Banesh Hoffmann and Leopold Infeld (hereafter referred to as EIH) made a fundamental advance in the problem of particle dynamics in general relativity. They were able to derive approximate equations of motion directly from the Einstein field equations for slowly moving, compact bodies without the need of additional assumptions. In the lowest, so-called Newtonian approximation, they obtained equations of motion identical in form to those of Newton for compact bodies interacting according to his inverse square law of gravitation. This result shows incidentally that general relativity encompasses all of the results of Newtonian mechanics. The EIH calculations also yielded directly the equivalence of inertial and gravitational masses.

Higher-order, post-Newtonian corrections to these laws correctly predict, among other observed effects, the perihelion advance of the planetary orbits. It is also possible to extend the EIH approximation method to obtain corrections that include the radiation reaction effects that have been observed in binary pulsar systems. Also, this author was able to derive the electromagnetic interaction force law between charged particles from the combined Einstein–Maxwell field equations.

In their original work EIH derived their equations of motion for bodies moving in an everywhere constant ($g_{\mu\nu} = \eta_{\mu\nu}$) external gravitational field. One can also use their approach to derive approximate equations of motion for compact bodies moving in external fields that are slowly varying in space and time. In particular one can derive an approximate version of the geodesic Eq. (19) for a single body, making it unnecessary to postulate the latter equations.

To better appreciate the significance of the EIH results it is useful to compare the problem of motion in general relativity to that in special relativity and Newtonian mechanics. In those theories one must postulate not only the force law between bodies, such as the Lorentz force law in electrodynamics, but also the form of the inertial terms in the equations of motion for these bodies. General relativity is unique among field theories in that both the force laws and inertial terms follow from the field equations of the theory.

D. Clocks, Rods, and Coordinates

It has been argued that some kind of postulate concerning the behavior of clocks and measuring rods is required in general relativity. For example, it has been suggested that a class of objects, ideal clocks, measure proper time along their trajectories, where the proper time along a trajectory is defined to be the integral of the distance ds, given by Eq. (22), along this trajectory. In this view, clocks, and also measuring rods, are assumed to be primitive objects in the theory.

Actually, all such postulates are unnecessary, in both the special and general theories. Clocks, and similarly measuring rods, are, in fact, composite physical systems with laws of motion governing their behavior. Once these laws have been established, there is no need to add additional postulates governing their behavior. It can be shown, for example, that if one takes, as a model for a clock, a classical hydrogen atom, then as long as the forces acting on this clock are small compared to the internal forces acting on its constituents and its dimensions are small compared to the curvature of its trajectory, it will indeed measure approximately the proper time along this trajectory. However, if the forces acting on it are sufficiently strong, the atom will be ionized and cease to measure any kind of time along its trajectory. Thus, the behavior of clocks is seen to be a dynamical question that cannot be decided *a priori* from any kinematic postulate.

In this view, then, clocks and measuring rods, and indeed all measuring devices, are considered to be physical systems with the geometrical objects associated with them obeying their own laws of motion. Furthermore, a physical description would have to be considered incomplete if it did not supply these laws of motion. To avoid the necessity of having to formulate and solve the laws of motion for a particular kind of clock, one may assume that it does satisfy the conditions for measuring proper time, with the proviso that if these conditions are violated it will no longer do so. If this assumption results in inconsistencies it does not mean that a principle of general relativity has been violated but only that these conditions have not been met.

While clocks and rods can be used to measure times and distances, it should be emphasized that these measurements bear no direct relation to the coordinates employed in the formulation of the laws of motion. Since, in all space–time descriptions, these laws are generally covariant, there are no preferred coordinate systems. Consequently, it follows that the predictions of a theory cannot depend on a particular coordinatization. In effect, the coordinates play the same role in space–time theories as do the indices that characterize the various components of a geometrical object and hence, like these indices, are not associated with any physical objects.

It is, however, often convenient to choose a particular coordinatization. Thus in Newtonian mechanics one usually chooses coordinates such that one of them is the parameter that characterizes the planes of simultaneity, and likewise in special relativity one usually employs inertial coordinates. In general relativity one also can employ a coordinatization that is particularly convenient for some purpose. One can, for example, choose coordinates in such a way that one of them corresponds to the time and distance intervals measured by a particular family of clocks and rods. Alternatively, one can choose coordinates so that certain components of the gravitational field have simple values. For example, one can choose coordinates so that $g_{00} = 1$ and $g_{01} = g_{02} = g_{03} = 0$. But in all cases such a choice is arbitrary and devoid of physical content.

V. GRAVITATIONAL FIELDS

A. Newtonian Fields

Since Newtonian theory describes, to a high degree of accuracy, the phenomena associated with weak gravitational fields, it is essential that this theory be an approximation to the general theory. Although originally formulated as an action-at-a-distance theory, the Newtonian theory of gravity can also be formulated as a field theory analogous to electrostatics. The gravitational field is characterized, in this version of the theory, by a single scalar field ϕ that satisfies, in suitable coordinates, the field equation

$$\nabla^2\phi = 4\pi G\rho, \tag{28}$$

where ρ is the mass density of the sources of the field and $G = 6.673$ cm^3 g^{-1} sec^{-2} is the Newtonian gravitational constant. For a point particle of mass m this equation has the well-known solution

$$\phi = -Gm/r. \tag{29}$$

In the general theory we assume that, in the case of weak fields, there exists a coordinate system such that $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$, where $h_{\mu\nu} \ll 1$. We also assume that the velocities of the sources of the gravitational field are all vanishingly small compared to the velocity of light. In this case, the only nonvanishing component of $T_{\mu\nu}$ is $T_{00} = \rho c^2$ and Eq. (25) can be shown to reduce to Eq. (28) if we set $\Lambda = 0$, $\kappa = -8\pi G/c^4$, where G is the Newtonian gravitational constant, and take

$$\phi = (c^2/2)h_{00}. \tag{30}$$

B. The "Flat" Field

The simplest exact solution of the empty-space Einstein equations with vanishing cosmological constant (Eq. (25) with $T_{\mu\nu} = 0$ and $\Lambda = 0$) is the so-called "flat" field. It satisfies Eq. (20), whence the appellation "flat," and in appropriate coordinates takes the form $g_{\mu\nu} = \eta_{\mu\nu}$. By itself this solution would be rather uninteresting if it were not for the fact that it is the only stationary solution that is asymptotically flat and everywhere regular (no singularities) on the whole space–time manifold $-\infty \le x^\mu \le \infty$ as was first shown by A. Lichnerowicz. One consequence of this result is that there do not exist regular, stationary solutions of the empty-space field equations that are asymptotically flat and that could be interpreted as "particle" solutions, that is, solutions that are nonflat only in some localized region of space. Einstein had hoped that such solutions of his field equations existed so that he could dispense with the matter stress–energy tensor $T^{\mu\nu}$ that appears in these equations.

C. The Schwarzschild Field, Event Horizons, and Black Holes

In spite of their enormous complexity, the Einstein field Eq. (25) possesses many exact solutions. One of the first and perhaps still the most important of these solutions was obtained by Schwarzschild in 1916 for the case $\Lambda = 0$ and $T_{\mu\nu} = 0$ by imposing the condition of spherical symmetry on $g_{\mu\nu}$. The nonvanishing components of $g_{\mu\nu}$ are given in spherical coordinates by

$$\begin{aligned} g_{00} &= 1 - 2M/r & \quad g_{11} &= -1/(1 - 2M/r) \\ g_{22} &= -r^2 & \quad g_{33} &= -r^2\sin^2\theta, \end{aligned} \tag{31}$$

where M is a constant of integration. This solution is seen to be independent of the coordinate x^0 and hence is a static field. The condition that the field be independent of x^0 was originally imposed by Schwarzschild in obtaining his solution of the field equations but has since been shown to be a consequence of the condition of spherical symmetry.

The importance of the Schwarzschild solution lies in the fact that it is the general relativistic analog of the Newtonian field (29) of a point mass. By making use of Eq. (30) and Eq. (31) for g_{00} we see that $M = Gm/c^2$. The constant $2M$ is refferred to as the Schwarzschild radius of the mass m. The Schwarzschild radius of the sun is 2.9 km and of the earth is 0.88 cm. For comparison, the Schwarzschild radius of a proton is 2.4×10^{-52} cm and that of a typical galaxy of mas $\sim 10^{45}$ g is $\sim 10^{17}$ cm.

The Schwarzschild field has a property that distinguishes it from the corresponding Newtonian field: at $r = 2M$ it becomes singular. Indeed, at this radius g_{11} is infinite! However, this is not a physical singularity, as Eddington first showed in 1924, but rather what is called a coordinate singularity. A final clarification of the structure of the Schwarzschild field came in 1960 with the work of M. Kruskal. He found a coordinate transformation from the Schwarzschild coordinates (x^0, r, θ, ϕ) to the set (u, v, θ, ϕ), where

$$u = a\cosh(x^0/4M) \qquad v = a\sinh(x^0/4M) \tag{32}$$

with $a = [(r/2M) - 1]^{1/2}\exp(r/4M)$, such that the transformed components of the Schwarzschild field were given by

$$g_{00} = f \qquad g_{11} = -f \qquad g_{22} = -r^2$$
$$g_{33} = -r^2\sin^2\phi, \tag{33}$$

where $f = (8M/r)\exp(-r/2M)$. In these latter expressions, r is a function of u and v obtained by solving Eq. (32) for this quantity. In this form the field is seen to be singular only at $r = 0$, which is a true physical singularity.

Figure 1 depicts a number of the features of the Kruskal transformation in what is called a Kruskal diagram. In this diagram, curves of constant Schwarzschild r correspond to the right hyperbolas $u^2 - v^2 = \text{const}$, while curves of constant x^0 correspond to straight lines passing through the origin. The region I to the right of the two $r = 2M$ lines corresponds to the entire region $r > 2M$, $-\infty < x^0 < \infty$. Hence it follows that the transformation from Kruskal to Schwarzschild coordinates must be a singular transformation, and indeed it is the singular nature of this

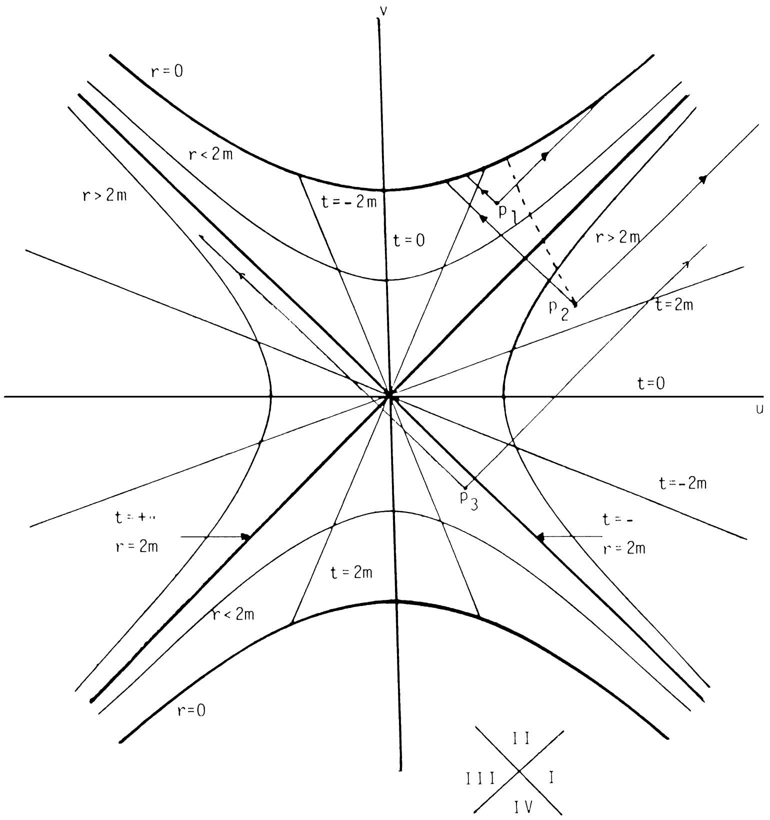


FIGURE 1 Kruskal diagram for the Schwartzschild field. The hyperbola $r = 0$ represents a real singularity of the field.

transformation that is responsible for the singularity of $g_{\mu\nu}$ at $r = 2M$ in Schwarzschild coordinates.

Many of the properties of the Schwarzschild field can be understood by referring to the Kruskal diagram. In Kruskal coordinates, light rays propagate along straight lines at 45° with respect to the u, v axes. As a consequence, it is seen that all of the rays emitted at a point P_1 in region II above the two lines $r = 2M$ will ultimately reach the singular curve $r = 0$, so no information can be transmitted from this point into region I. Likewise, inward-directed rays emitted at a point P_2 in region I will also reach the $r = 0$ curve while outward-directed rays will continue their outward propagation forever. And finally, some of the rays emitted from a point P_3 in region IV below the two lines $r = 2M$ will ultimately reach points in region I while others will reach the region III to the left of the two lines $r = 2M$.

Because of these features, the surface $r = 2M$ is said to form an event horizon: an observer in region I can never receive information about events taking place in regions II and III of the diagram. In addition, it is seen to be a light cone.

Finally, we note that a material body starting from rest at P_2 will fall into the singularity at $r = 0$ along the dashed path indicated in Fig. 1. Even though it reaches the event horizon at $x^0 = \infty$, the proper time along the curve from P_2 to the point where it reaches the singularity is finite. A local observer falling with the particle would notice nothing peculiar as he passed through the horizon. As long as the particle is outside this horizon it is possible to reverse its motion so that it does not cross it. However, once it does its fate is sealed; its motion can no longer be reversed and it will ultimately reach the $r = 0$ singularity in a finite amount of proper time. For this reason the source of a Schwarzschild field is said to be a black hole—nothing that falls into it can ever get out again and any radiation generated inside the horizon can never be seen from the outside.

D. Kerr–Newman Fields, Naked Singularities

In addition to the Schwarzschild family of solutions, each member of which is characterized by a value of the parameter M, other stationary solutions of the empty-space Einstein field equations have been found by Kerr and Newman. Like the Schwarzschild solution, the Kerr–Newman solutions are asymptotically flat; that is, the value of the Riemann–Christoffel tensor goes to zero as the coordinate r approaches infinity and also the physical singularity in the solution is surrounded by an event horizon. This family of solutions is characterized by three continuously variable parameters, which, because of the asymptotic form of these solutions, can be interpreted as the mass, angular momentum, and electric charge of the black hole.

After the discovery of the Kerr–Newman solutions it was shown by the work of Israel, Carter, Hawking, and Robinson that the Kerr–Newman solutions are the only asymptotically flat, stationary black hole solutions, that is, solutions with event horizons that depend continuously on a finite set of parameters. This result is a version of a conjecture of Wheeler to the effect that "a black hole has no hair." What is still lacking is a proof of uniqueness of the Kerr–Newman solutions, that is, that not more than three parameters is sufficient to characterize completely all stationary, asymptotically flat black hole solutions. To date, the best one can do in this respect is to show that the only such static solutions and the only such neutral solutions belong to the family of Kerr–Newman solutions.

In addition to the Kerr–Newman family of solutions there are solutions of the empty-space field equations that do not have event horizons surrounding the physical singularity in the solution. One such solution is obtained by letting the parameter M in the Schwarzschild solution become negative. Also, a charged Schwarzschild field has no horizon if the charge $Q < M$. In both cases the only singularity occurs at $r = 0$. Because of the absence of a horizon surrounding this point, it is referred to as a naked singularity.

E. Fields with Matter—Gravitational Collapse

In addition to the empty-space solutions discussed here there are numerous solutions of the Einstein equations with nonvanishing energy–momentum tensors. One can, for example, construct a spherically symmetric, nonsingular interior solution that joins on smoothly to an exterior Schwarzschild solution. The combined solution would then correspond to the field of a normal star, a white dwarf, or a neutron star. What distinguishes these objects is the equation of state for the matter comprising them. It is surprising that no interior solutions that could be joined to a Kerr–Newman field have been found.

Normal stars, of course, are not eternal objects. They are supported against gravitational collapse by pressure forces whose source is the thermonuclear burning that takes place at the center of the star. Once such burning ceases due to the depletion of its nuclear fuel, a star will begin to contract. If its total mass is less than approximately two solar masses it will ultimately become a stable white dwarf or a neutron star, supported by either electron or neutron degeneracy pressure. If, however, its total mass exceeds this limit, the star will continue to contract under its own gravity down to a point, a result first demonstrated by Oppenheimer and Snyder in 1939. Although they treated only cold matter, that is, matter without pressure, their result would still hold once the star has passed a certain critical stage even if the repulsion of the nuclei comprising the star were infinite. This critical stage is reached once the radius of the star becomes less than its Schwarzschild radius. When this happens, the surface $r = 2M$ becomes an event horizon and the matter is trapped inside, forming a black hole. Even an infinite pressure would then be unable to halt the continued contraction to a point because such a pressure would contribute an infinite amount to the energy–momentum tensor of the matter, which in turn would result in an even stronger gravitational attraction.

Because of the disquieting features of black hole formation (neither Eddington nor Einstein was prepared to accept their existence), theorists looked for ways to avoid their formation. However, theorems of Penrose and Hawking show that collapse to a singularity is inevitable once the gravitational field becomes strong enough to drag back any light emitted by the star, that is, when the escape velocity at the surface of the star exceeds the velocity of light.

What has not been proved to date is what Penrose calls the hypothesis of "cosmic censorship." This hypothesis asserts that matter will never collapse to a naked singularity, but rather that the singularity will always be surrounded by an event horizon and hence not be visible to an external observer. While the hypothesis is suported by both numerical and perturbation calculations, it has so far not been shown to be a rigorous consequence of the laws of motion of the general theory.

The detection of black holes is complicated by the fact that they are black—by themselves they can emit no radiation. If they exist at all then, they can be detected only through the effects of their gravitational field on nearby matter. If a black hole were a member of a double star system, it would become the source of intense X rays when its companion expanded during the later stages of its own evolution. As matter from the companion fell onto the black hole it would become compressed and thus heated to temperatures high enough for it to emit such high-energy radiation. Of course, it is not enough to find an x-ray-emitting binary system in order to prove the existence of a black hole. It is also necessary that the mass of the x-ray-emitting component be larger than the upper limit on the mass of a stable neutron star, that is, larger than 3 $M_\odot$ ($M_\odot$ denotes a solar mass of 1.99×10^{30} kg).

The first object to be definitely identified as a black hole in 1971 was a member of a binary system Cygnus X-1 in the constellation Cygnus which was an intense emitter of x-rays. Measurements established that the mass of the compact component of the system was about 8 $M_\odot$. Since then eight more black holes have been found in binary systems. In addition to stellar mass black holes there is now compelling evidence that massive black holes exist in the centers of galaxies. By measurements of the orbital speed of the gaseous disk around the nucleus of the spiral galaxy NGC 4258, Makoto Miyoshi and his collaborators were able to establish in 1995 that the mass of the central object is 4×10^7 $M_\odot$ and its diameter is a half a light year, thereby establishing its identity as a black hole. Our own galaxy has been shown to contain a black hole at its center of mass 2.6×10^6 $M_\odot$. To date, there are about 15 masses that have been determined for black holes in the centers of nearby galaxies. The most massive such candidate for a black hole is in the giant elliptic galaxy M87 with an estimated mass of 3×10^9 $M_\odot$ although it is not certain that this object is indeed a black hole. Finally it is now believed that the energy source of quasars is the inflow of vast amounts of gas into massive black holes in the centers of very young galaxies.

F. Gravitational Radiation, the Quadrupole Formula

Shortly after he formulated the general theory, Einstein showed that, when linearized around the flat field $g_{\mu\nu} = \eta_{\mu\nu}$, the empty-space field equations with zero cosmological constant possessed plane wave solutions similar to the plane wave solutions of electromagnetism. These waves propagate along the light cones defined by $\eta_{\mu\nu}$, that is, at the speed of light. Like their electromagnetic counterparts, they are transverse waves and possess two independent states of polarization. However, unlike the dipole structure of plane electromagnetic waves, gravitational waves have a quadrupole structure, as would be expected from the tensor nature of the gravitational field. Using coordinates such that $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$ and the direction of propagation as the z axis, the two states of polarization are characterized by $h^{+}_{xx} = h^{+}_{yy}$ and $h^{\times}_{xy} = h^{\times}_{yx}$ with, in the two cases, the other components having the value zero.

Although the empty-space field equations have been shown to possess exact solutions with wavelike properties, the so-called plane fronted waves, neither they nor the approximate wave solutions found by Einstein are associated with sources. In electromagnetic theory it is possible to find exact solutions of the field equations associated with sources with arbitrary motions. So far, no such solutions have been found for the highly nonlinear equations of the general theory. In order to construct radiative solutions associated with sources it is necessary to employ approximation methods.

The first attempts to carry out such approximations were made by Einstein and Arthur Eddington. Einstein tried to calculate the energy radiated by a localized system of masses by introducing a pseudo-energy-momentum tensor for the gravitational field. However, this object was not a true tensor and in fact it was later recognized that it was in general not possible to introduce a true energy-momentum tensor for the gravitational field. As a consequence, even if its components were all zero in one coordinate system they could be non-zero in some other system of coordinates. Using radiative solutions of the linearized gravitational field equations similar to the Liénard–Wiechert solutions of Maxwell's equations for an arbitrarily moving point charge, he derived the now famous quadrupole formula for the "energy" emitted by these masses. The derivation however was unsatisfactory because of its use of a pseudotensor to calculate the energy radiated by the system and the assumption that this "energy" loss was equal to the change in the energy of the radiating system. Eddington, on the other hand worked directly with the near field of a spinning rod held together with cohesive forces, which were assumed to be large compared to the gravitational forces between parts of the rod. He also used the linearized equations to calculate this field and used it together with the conservation laws derivable from the field equations to calculate the force exerted by one part of the rod on another. He was thus able to calculate the energy loss by the rod without having to make use of Einstein's pseudotensor. The result he obtained for the rate of energy loss, $\frac{32}{5}kI^2\omega^6$, where $k = 2.7 \times 10^{-60}$ for cgs units, ω is the angular frequency of the rod and I is its moment of inertia about the axis of rotation, agreed with the result obtained by Einstein. Eddington was careful to point out that this result could not be applied directly to systems such as binary stars held together by gravitational forces because in such systems non-linear gravitational effects can not be ignored.

After these initial derivations of Einstein and Eddington, much work went into a satisfactory derivation of the effects of gravitational radiation on gravitationally bound systems. In fact, for a time Einstein and others believed that real gravitational radiation did not exist. Because of the need to take account of the non-linear terms in the field equations many different approximation procedures were employed for this purpose. Today the best that can be done is to use an extension of the EIH procedure using matched asymptotic expansions and multiple time scale methods to calculate these effects. As a consequence these results can only be applied to systems whose component velocities are small compared to the speed of light and for which the separations between the components are large compared to their size. The net result of these calculations is the so-called quadrupole radiation formula

$$dE/dt = -\tfrac{1}{5}(G/c^5)\big\langle d^3Q_{ij}\big/dt^3 d^3Q_{ij}\big/dt^3\big\rangle, \qquad (34)$$

where the angle brackets denote an average over a period of oscillation of the source and the indices i and j take on the values 1 through 3. The quantities Q_{ij} are the components of the mass quadrupole moment:

$$Q_{ij} = \int \rho(x)\left(x_i x_j - \frac{1}{3}\delta_{ij}\delta^{kl}x_k x_l\right)dv, \qquad (35)$$

where $\rho(x)$ is the mass density of the source. Finally, the quantity E appearing on the left side of this equation is the total Newtonian energy, kinetic plus potential, of the source. As a consequence, the quadrupole formula can be interpreted as an expression of energy conservation, in which case the quantity on its right side becomes the energy carried away by gravitational waves.

The quadrupole formula gives the main contribution to the energy loss of a system due to gravitational radiation. Changes in higher multipoles of the mass will also contribute to this loss, although in general their contribution will be much smaller than that of the quadrupole. However, unlike the electromagnetic case, there is no dipole contribution, since both the total momentum and angular momentum of the radiating system are conserved. And in both theories there is no monopole radiation because of conservation of charge in the one case and conservation of mass in the other.

The amount of energy emitted by slow-motion sources is in most cases very small. A beryllium rod of length 170 m and weighing 6×10^7 kg, spinning on an axis through its center as fast as it can without disintegrating ($\omega \sim 10^3$ sec^{-1}), would radiate energy at the rate of approximately 10^{-7} erg/sec. For the earth–sun system Eq. (34) yields a rate of energy loss of about 200 W. Only in the case of extremely massive objects moving at high speeds such as in the binary pulsar will the amount of energy radiated be significant.

VI. OBSERVATIONAL TESTS OF GENERAL RELATIVITY

A. Gravitational Red Shift

In general relativity the apparent rate at which clocks run is affected by the presence of a gravitational field. Like its counterpart in special relativity, this is a kinematic effect and hence is independent of any direct effect of the gravitational field on the internal dynamics of the clock. Only if gradients of this field result in tidal forces that are comparable to the nongravitational forces responsible for the functioning of the clock will this internal dynamics be altered.

The amount of red shift is most easily calculated in the case of a static gravitational field and for two clocks that are at rest in this field. We suppose that one clock, the emitter, sends out light waves whose frequency ν_{em} is the same as its own frequency. A receiving clock, which is identical in construction to the emitting clock, that is, has the same internal dynamics, is used to measure the frequency ν_{rec} of the received radiation. Then it can be shown that

$$Z := (\nu_{\text{rec}} - \nu_{\text{em}})/\nu_{\text{em}} = (1/c^2)(\phi_{\text{em}} - \phi_{\text{rec}}), \tag{36}$$

where ϕ_{em} and ϕ_{rec} are the gravitational potentials at the locations of the emitter and receiver, respectively. If ϕ_{em} is less than ϕ_{rec} then the quantity Z is negative, hence the emitted light appears to be red shifted. This effect can be understood by noting that, in going from the emiter to the receiver, a photon will, in this case, gain potential energy. Since its total energy is conserved, its kinetic energy, which is proportional to its frequency, will decrease and hence so will its frequency. If either the emitter or receiver is moving with respect to the background field then Eq. (36) must be amended to take account of the Doppler shift produced by such motion.

The first attempts to observe the gravitational red shift were made on the spectral lines of the sun and known white dwarfs. For the sun, $Z = -2.12 \times 10^{-6}$ while for white dwarfs it would have values 10–100 times as large. In the case of the sun the shift was masked by the Doppler broadening of the spectral lines due to thermal motion. However, observations near its edge are consistent with a red shift of the magnitude predicted by Eq. (36). In the case of the white dwarfs, it was not possible to measure their masses and radii with sufficient accuracy to determine ϕ_{em}, although again red shifts were observed whose magnitudes were consistent with estimates of these quantities.

The first accurate test of the red shift prediction was carried out in a series of terrestrial experiments by Pound and Rebka in 1960, using the Mössbauer effect. The emitter and detector used by them were separated by a vertical height of 74 ft. In this case the gravitational potential can be taken equal to gz, where g is the local acceleration due to gravity and z is the height above ground level. Equation (36) then yields the value $Z = 2.5 \times 10^{-15}$. In spite of its small value, Pound and Rebka were able to observe a shift equal to 1.05 ± 0.10 times the predicted Z. Later experiments with cesium beam and rubidium clocks on jet aircraft yielded similar results.

The possibility of testing the red shift prediction has improved dramatically with the development of high-precision clocks. In 1976 Vessot and Levine used a rocket to carry a hydrogenmaser clock to an altitude of about 10,000 km. Their result verified the theoretical value to within 2 parts in 10^{-4}.

It has been argued that red shift observations do not bear on the question of the validity of general relativity but rather only on the validity of the principle of equivalence. This is, in fact, only partially true. If one assumes that the gravitational field of the earth is uniform over the 74 ft that separated the emitter and receiver of the Pound–Rebka experiment, then indeed their result can be calculated using only this principle. However, one cannot use it alone to determine the result of the Vessot–Levine experiment. In this case it is necessary to make direct use of Eq. (36). The derivation of Eq. (36), however, makes use of Eq. (19) for a light ray, which, in turn, is a consequence of the field equations (25) of general relativity. Also, although the present red shift measurements are not sufficiently accurate to distinguish between different possible equations for the gravitational field, there is nothing in principle that would preclude such a test.

B. Solar System Tests—The PPN Formalism

The first arena used for testing general relativity was the solar system and it remains so to this date. What has changed dramatically over the years, due to the rapid growth of technology, is the degree of accuracy with which the theory can be tested. It is in the solar system that the gravitational field of the sun is of sufficient strength that deviations from the Newtonian theory are observable, but just barely.

In calculating the size of these effects one assumes that the trajectories of planets and light rays obey Eqs. (19). However, rather than take the sun's gravitational field to be the Schwarzschild field in evaluating the Christoffel symbols appearing in these equations, it is useful to use what has become known as the parametrized post-Newtonian (PPN) formalism. In this formalism, first developed by Eddington and extended by Robertson, Schiff, Will, and others, one assumes a more general form for the post-Newtonian corrections to the gravitational field of the sun than those given by general relativity. These corrections are allowed to depend on a number of unknown parameters that one hopes to determine by solar system and other observations. The reason for proceeding in this manner is that it allows one to test the validity of other competing theories of gravity in which these parameters have different values from those they would have if general relativity were valid.

In the most extreme versions of this formalism as many as 10 parameters are employed. However, a number of these parameters can be eliminated if one requires that the equation of motion (19) are a consequence of the field equations for the gravitational field, as they are in general relativity. Also, some of these parameters are known to be small from other experiments. In what follows we will use an abbreviated version of the PPN formalism employed by Hellings in his analysis of the solar system data. In this version the components of the gravitational field have the form

$$\begin{aligned} g_{00} &= 1 - 2U\big[1 - J_2(R_\odot/r)^2 P_2(\theta)\big] \\ &\quad + 2\beta U^2 + \alpha_1 U(w/c) && \text{(37a)} \\ g_{0i} &= \alpha_1 U w^i\big/c && \text{(37b)} \\ g_{ij} &= -(1 + 2\gamma U)\delta_{ij}, && \text{(37c)} \end{aligned}$$

where β, γ, and α_1 are PPN parameters $U = GM_\odot/rc^2$ is the Newtonian gravitational potential of the sun, and $R_\odot$ and $M_\odot$ are the radius and mass of the sun. Included in Eq. (37a) is a term proportional to J_2, a dimensionless measure of the quadrupole moment of the sun. In this term P_2 is the Legendre polynomial of order 2 and θ is the angle between the radius vector from the sun's center and the normal to the sun's equator. The so-called preferred frame velocity w^i is taken to be the average of the determinations of the solar system velocity relative to the cosmic blackbody background. In general relativity $\beta = \gamma = 1$ and $\alpha_1 = 0$.

1. Bending of Light

One of the more spectacular confirmations of the general theory came in 1919, when the solar eclipse expedition headed by Eddington announced that they had observed bending of light from stars as they passed near the edge of the sun that was in agreement with the prediction of the theory. Derivations of the bending that use only the principle of equivalence or the corpuscular theory of light predict just half of the bending predicted by the general theory.

The angle of bending for light passing at a distance d from the center of the sun can be computed by using the equation of motion (19) with $g_{\mu\nu}\, dx^\mu/d\lambda\ dx^\nu/d\lambda = 0$. Using the form for the gravitational field given by Eq. (37), the angle of bending is given by

$$\Theta = (1+\gamma)2GM_\odot\big/c^2 d. \tag{38}$$

For light that just grazes the edge of the sun $\Theta = 1''.75$ when $\gamma = 1$. Because of this small value it is necessary to observe stars whose light passes very close to the edge of the sun, and this can be done only during a total eclipse. The apparent positions of these stars during the eclipse are then compared to their positions when the sun is no longer in the field of view in order to measure the amount of bending. Unfortunately, such measurements are beset with a number of uncertainties. Thus the measurements made by Eddington and his co-workers had only 30% accuracy. The most recent such measurements were made during the solar eclipse of June 30, 1973, and yielded the value

$$\tfrac{1}{2}(1+\gamma) = 0.95 \pm 0.11. \tag{39}$$

The use of long-baseline and very-long-baseline interferometry, which is capable in principle of measuring angular separations and changes in angle as small as 3×10^{-4}, has made possible much more accurate tests of the bending of light. These techniques have been used to observe a number of quasars such as 3C273 that pass very close to the sun in the course of a year. Beginning in 1970, these observations have yielded increasingly accurate determinations, and the most recent, in 1984, agrees with the general relativistic prediction to within 1%.

2. Time Delay

In passing through a strong gravitational field, light not only will be red shifted but also will take longer to traverse a given distance than it would if Newtonian theory were valid. The reason for this delay is that the gravitational field acts like a variable index of refraction, so the velocity of light will vary as it passes through such a field. This effect was first proposed by Shapiro in 1964 as a means of testing general relativity. It can be observed by bouncing a radar signal off a planet or artificial satellite and measuring its round-trip travel time. At superior conjunction, when the planet or satellite is on the far side of the sun from the earth, the effect is a maximum, in which case the amount of delay is given by

$$\delta t = 2(1+\gamma)\big(GM_\odot\big/c^3\big)\ln\big(4r_{\rm e}r_{\rm p}\big/d^2\big), \tag{40}$$

where $r_{\rm c}$, $r_{\rm p}$, and d are, respectively, the distance from the sun to the earth, the distance from the sun to the target, and the distance of closest approach of the signal to the center of the sun. Since one does not have a Newtonian value for the round-trip travel time with which to compare the measured time it is necessary to monitor the travel time as the target passes through superior conjunction and look for a logarithmic dependence.

The use of a planet such as Mercury or Venus as a target is complicated by the fact that its topography is largely unknown. As a consequence, a signal could be reflected from a valley or a mountaintop without our being able to detect the difference. Such differences can introduce errors of as much as 5 μsec in the round-trip travel time. Artificial satellites such as Mariners 6 and 7 have been used to overcome this difficulty. Furthermore, since they are active retransmitters of the radar signal they permit an accurate determination of their true range. Unfortunately, fluctuations in the solar wind and solar radiation pressure produce random accelerations that can lead to uncertainties of up to 0.1 μsec in the travel time. Finally, spacecraft such as the Mariner 9 Mars orbiter and the Viking Mars landers and orbiters have been used as targets. Since they are anchored to the planet they will not suffer such accelerations. The most recent measurements by Reasenberg *et al.* in 1979 have yielded a value

$$(1+\gamma)/2 = 1.000 \pm 0.001. \tag{41}$$

3. Planetary Motion

Long before the general theory was proposed, it was known that there was an anomalous precession of the perihelion (distance of closest approach to the sun) of the planet Mercury that could not be accounted for on the basis of Newtonian theory by taking into consideration the perturbations on Mercury's orbit due to the other planets. At the end of the last century, Newcomb calculated this residual advance to have a value of $41''.24 \pm 2''.09$ of arc per century.

The field values given by Eq. (37) and the equation of motion (19) together yield an expression for the perihelion advance per period that is given, to an accuracy commensurate with the accuracy of the observations, by

$$\delta\bar{\omega} = (6\pi GM_\odot/c2p)\big[\tfrac{1}{3}(2+2\gamma-\beta) + J_2\big(R_\odot^2 c^2\big/12GM_\odot p\big)\big], \qquad (42)$$

where $p = a(1-e^2)$ is the semi-latus rectum of the orbit, a its semimajor axis, and e its eccentricity. Using the best current values for the orbital elements and physical constants for Mercury and the sun, one obtains from Eq. (38) a perihelion advance of $42''.95\lambda_p$ of arc per century, where $\lambda_p = [\frac{1}{3}(2+2\gamma-\beta) + 3\times 10^3 J_2]$.

The measured value of the perihelion advance of Mercury is known to a precision of about 1% from optical measurements made over the past three centuries and of about 0.5% from radar observations made over the past two decades. If one assumes that J_2 has the value $\sim 1\times 10^{-7}$, which it would have if it were the consequence of centrifugal flattening due to a uniform rotation of the sun equal to its observed surface rate of rotation, then, using this value, Shapiro gives

$$\tfrac{1}{3}(2+2\gamma-\beta) = 1.003 \pm 0.005, \qquad (43)$$

which is in excellent agreement with the prediction of general relativity.

This agreement has been called into question by some researchers, notably Dicke and Hill. Observations of the solar oblateness by Dicke and Goldenberg in 1966 led them to conclude that J_2 actually has a value of $(2.47 \pm 0.23)\times 10^{-5}$, leading to a contribution of about $4''$ per century to the overall perihelion advance. If true, this would put the prediction of general relativity into serious disagreement with the observations. On the other hand, it would agree with the prediction of the Brans–Dicke scalar tensor theory of gravity if an adjustable parameter in that theory were suitably chosen. However, a number of authors have disagreed with the interpretation of their observations by Dicke and Goldenberg. These authors argue that the observations could equally well be explained by assuming a standard solar model with $J_2 \sim 10^{-7}$ and a surface temperature difference of about 1° between the pole and the equator. More recently Hill has given a value of $J_2 = 6\times 10^{-6}$, based on his measurements of normal mode oscillations of the sun. If true, the general relativistic prediction for Mercury would be inconsistent with the observed value by about two standard deviations. Unfortunately, the present measurements of the orbit of Mercury are not sufficiently accurate to separate the post-Newtonian and quadrupole effects.

A resolution of this difficulty has come from an analysis of the ranging data for the planet Mars. Since the quadrupole contribution to the perihelion advance has a different dependence on the semimajor axis from the gravitational effect, it is in principle possible to separate the two by observing the advance for different planets. In spite of the smallness of these effects on the orbit of Mars, the accuracy of the Viking data from Mars, which are accurate to within 7 km, combined with the radar data from Mercury allows such a determination. Using a solar system model that includes 200 of the largest asteroids. Hellings has found, with $J_2 = 0$, that

$$\beta - 1 = (-0.2 \pm 1.0)\times 10^{-3} \qquad (44a)$$

$$\gamma - 1 = (-1.2 \pm 1.6)\times 10^{-3} \qquad (44b)$$

$$\alpha_1 = (2.2 \pm 1.8)\times 10^{-4}. \qquad (44c)$$

When J_2 was allowed to have a finite value, he found that

$$J_2 = (-1.4 \pm 1.5)\times 10^{-6} \qquad (45a)$$

and

$$\beta - 1 = (-2.9 \pm 3.1)\times 10^{-3} \qquad (45b)$$

$$\gamma - 1 = (-0.7 \pm 1.7)\times 10^{-3} \qquad (45c)$$

$$\alpha_1 = (2.1 \pm 1.9)\times 10^{-4}. \qquad (45d)$$

Hellings also used these data to analyze the nonsymmetric gravitational theory of Moffat, which was consistent with the Mercury data and Hill's value for J_2. The result was that

$$J_2 = (1.7 \pm 2.4)\times 10^{-7}. \qquad (46)$$

From these results it appears that the predictions of general relativity are confirmed to about 0.1%. However, by a suitable adjustment of parameters, several competing theories also share this property. What distinguishes general relativity from these other theories is that, aside from the value for the gravitational constant G, it contains no other adjustable parameters.

4. Time Varying G

In addition to the tests discussed above, the solar system data can be used to test the possibility that the gravitational constant varies with time. Such a possibility was first suggested by Dirac in 1937 on the basis of his large number hypothesis. He observed that one could form, from the atomic and cosmological constants, several dimensionless numbers whose values were all of the order of 10^{40}. Rather than being a coincidence that was valid only at the present time, Dirac proposed that the equality of these numbers was the manifestation of some underlying physical principle and that they held at all times. Since one of these numbers involves the present age of the universe through its dependence on the Hubble "constant" and hence decreases as one moves back in time, the other constants must also change with time in order to maintain the equality between the large numbers. One of these numbers, however, involves only atomic constants, being the ratio of the electrical to the gravitational force between an electron and a proton. Hence the Dirac hypothesis requires that one of these atomic constants must be changing on a cosmic time scale. The constant that is usually taken to vary with time in theoretical implementations of the large number hypothesis is the gravitational constant.

There are several ways of constructing a theory with an effective time-varying gravitational constant. In the Brans–Dicke theory, the effective gravitational constant itself varies with time:

$$G_{\text{eff}} = G[1 + (\dot{G}/G)(t - t_0)]. \tag{47}$$

An alternative proposal by Dirac assumed that cosmic effects couple to local atomic physics so that the ratio of atomic to gravitational time is not constant. The rate of change of gravitational time τ_{G} with respect to atomic time τ_{A} is then given as

$$d\tau_{\text{G}}/d\tau_{\text{A}} = 1 + \dot{\phi}(t - t_0), \tag{48}$$

where ϕ is some cosmological field that is supposed to be responsible for the effect. In both cases, the net effect is to produce an anomalous acceleration in the equations of motion for material bodies.

Since the change in atomic constants is tied to cosmic evolution in the large number hypothesis, the expected rate of change in G should be proportional to the inverse Hubble time:

$$\dot{G}/G - H_0 \cong 5 \times 10^{-11}\ \text{year}^{-1} \tag{49}$$

On the basis of the Viking lander data, Hellings concludes that

$$\dot{G}/G = (0.2 \pm 0.4) \times 10^{-11}\ \text{year}^{-1} \tag{50a}$$

$$\dot{\phi} = (0.1 \pm 0.8) \times 10^{-11}\ \text{year}^{-1} \tag{50b}$$

Since these limits are an order of magnitude smaller than what one would expect from simple cosmic scale arguments, they cast serious doubt on the large number hypothesis.

C. The Binary Pulsar

A new, and essentially unique, opportunity for testing general relativity came with the discovery of the binary pulsar PSR 1913 + 16 by Hulse and Taylor in 1974. It consists of a pulsar in orbital motion about an unseen companion with a period of 7.75 hr. Its relevance for general relativity is twofold: because $v^2/c^2 \sim 5 \times 10^{-7}$ is a factor 10 larger than for Mercury, relativistic effects are considerably larger than any that have been observed in the solar system. Also, the short period amplifies secular changes in the orbit. Thus the observed periastron advance amounts to $4^\circ.2261 \pm 0.0007$ of arc per year compared to the $43''$ of arc per century for Mercury. Furthermore, the pulsar carries its own clock with a period that is accurate to better than one part in 10^{12}. As a consequence, measurements of post-Newtonian effects can be made with unprecedented accuracy. If this were all, the binary pulsar would still be an invaluable tool for testing general relativistic orbit effects. However, it also provides us for the first time with a means for testing an essentially different kind of prediction of general relativity, namely the existence of gravitational radiation.

Considerable effort has gone into identifying the pulsar companion. It was soon found that the pulsar radio signals were never eclipsed by the companion. Also, the dispersion of the pulsed signal showed little change over an orbit, implying the absence of a dense plasma in the system. These two facts together ruled out the possibility of the companion being a main sequence star. Another possibility is that it is a helium star. However, since the pulsar is at a distance of only about 5 kpc from us, such a star would have been seen. In spite of intense efforts, no such star has been observed in the neighborhood of the pulsar. The remaining possibility is that it is a compact object, either a white dwarf, another neutron star, or a black hole.

In the case of conventional spectroscopic binaries, it is usually possible to measure only two parameters of the system, the so-called mass function of the two masses M_1 and M_2 of the components and the product of the semimajor axis a_1 and the sine of the angle i of inclination of the plane of the orbit to the line of sight. However, in the case of the binary pulsar one can use general relativity to determine all four of these parameters from measurements of the periastron advance and the combined second-order Doppler shift and gravitational red shift of the emitted signals. One finds from these combined measurements that $M_1 = M_2 = (1.41 \pm 0.06)\ M_\odot$. From the fact that the Chandrasekhar limit on the mass of a nonrotating white dwarf is about 1.4 $M_\odot$, it appears likely that the unseen companion is either a neutron star or a black hole.

In addition to the measurements discussed above, it was discovered that the orbital period P was decreasing with time. Later measurements gave a value for $\dot{P} = (-2.30 \pm 0.22) \times 10^{-12}$ sec/sec^{-1} or about 7×10^{-5} sec/year. If one computes the period change due to loss of energy by the emission of gravitational radiation using the quadrupole formula (34) one obtains a value for $\dot{P} = -2.40 \times 10^{-12}$ sec/sec^{-1}, in excellent agreement with the observed value.

Of course, there are other effects that could change the orbital period, such as tidal dissipation, mass loss or accretion onto the system, or acceleration relative to the solar system. Furthermore, there could be other contributions to the periastron advance such as rotational or tidal deformation of the companion. Only in the case of a helium-star companion would any of these effects contribute significantly to the calculated or observed period change. Furthermore, it would be truly remarkable if some combination of these effects should conspire to give a value for the period change equal to that predicted by the quadrupole formula. It therefore appears safe to say that for the first time we have evidence of a qualitatively new prediction of general relativity, namely gravitational radiation. Finally, the data from the binary pulsar seem to rule out a number of competing theories of gravity such as the Rosen bimetric theory. In such theories this system can radiate dipole gravitational waves that result in a period increase. Only a very artificial mechanism could then give rise to the observed period decrease.

D. Gravitational Wave Detection

The first comprehensive attempt to detect gravitational waves impinging on the earth was begun by J. Weber in 1961. His antenna consisted of a large aluminum cylinder ~1.5 m long with a resonant frequency of ~1660 Hz. In the early 1970s Weber announced the detection of coincident pulses on two of these antennae separated by a distance of ~1000 mile. However, attempts to duplicate these results by a number of other groups, using somewhat more sensitive detectors than those used by Weber, proved fruitless, and it is now generally agreed that the events recorded by Weber were not caused by gravitational waves.

Since Weber's pioneering efforts, about 15 different groups from around the world have undertaken the construction of gravitational wave detectors. The sensitivity of a detector can be expressed in terms of the smallest strain $\Delta L/L$, where ΔL is a change in the length L of the detector, that can just be measured. This change in length is produced by tidal forces associated with the incident gravitational wave and hence its measurement leads to a determination of the Riemann–Christoffel tensor of the wave. Since this strain is approximately equal to the dimensionless amplitude h of a gravitational wave incident on the detector, the sensitivity of a detector is usually given as the minimum value of h that can be detected.

The original Weber bars had a sensitivity $h \sim 10^{-16}$. At present one of the main limitations on the sensitivity of Weber bars is thermal noise. As a consequence, second-generation Weber bars are being constructed that will be cooled to liquid helium temperatures. Such bars are estimated to have sensitivities of $h \sim 10^{-19}$. It is technically feasible to construct bars for which $h \sim 10^{-21}$, although that would require cooling to the millidegree level. The latter value appears to be a lower limit to what can be attained with presently available technology.

One of the drawbacks of Weber bar detectors is that they are only sensitive to the Fourier component of the incoming signal whose frequency is equal to the resonant frequency of the bar. Furthermore, most bars have resonant frequencies in the kiloherts range with a smallest reported frequency of 60.2 Hz. Unfortunately, most continuous wave sources such as binary star systems have much lower frequencies. In an attempt to overcome this difficulty and to increase sensitivity, a number of groups have undertaken the construction of laser interferometer detectors. In these devices, a gravitational wave would change the lengths of the interferometer arms and one would measure the resulting fringe shifts. Such detectors can, in principle, record the entire waveform of an incoming wave rather than a single Fourier component. It is possible that sensitivities as low as $h \sim 10^{-22}$ might be achieved. The most ambitious of these projects to date is the LIGO (for laser interferometer gravitational wave observatory) project which is building two such detectors in the United States and is expected to go on line in the next few years. It has also been suggested that gravitational radiation could be detected by the accurate Doppler tracking of spacecraft. Such a scheme would, in principle, be able to detect waves with frequencies in the 1 to 10^{-4} range. Present technology is within one or two orders of magnitude of the sensitivity needed to detect possible signals in this frequency range.

Possible sources of gravitational waves can be divided into two groups, those that emit continuously and those that emit in bursts. Possible continuous wave sources are binary stars and vibrating or rotating stars. In the case of binary stars, the strongest emitter known is μ Scorpii, for which $h = 2.1 \times 10^{-20}$. However, its frequency is 1.6×10^{-5} Hz. The largest binary frequency known is 1.9×10^{-3} Hz. However, for this system $h \sim 5 \times 10^{-22}$. Other possible continuous wave sources have values for h that are this small or smaller. Thus, estimates of h for waves from the Crab and Vela pulsars are of order 10^{-24} to 10^{-27}, at frequencies between 10 and 100 Hz. In spite of these low amplitudes, signal integration over an extended time can effectively increase the sensitivity of a detector by an order of magnitude or more, so the detection of such signals is not totally out of the question.

Bursts of gravitational radiation can be expected to accompany cataclysmic events such as supernova explosions, stellar collapse to form neutron stars or black holes, or coalescence of the neutron stars or black holes in a binary system at the end stage of its evolution. One of the problems in dealing with such systems is the determination of the efficiency with which other forms of energy can be converted into gravitational radiation. Estimates range from a maximum of 0.5 to as low as 0.001. Such events would have characteristic frequencies in the range 10^2 to 10^5 Hz and those occurring in our galaxy would have amplitudes estimated to be in the range $h \sim 10^{-18}$ to 10^{-17}. Here the problem for detection is not so much the frequency or the intensity, as it is in the case of continuous emitters, but rather the scarcity of such events. Thus, the supernova rate in our galaxy has been estimated to be 0.03 per year. If one includes such events in other galaxies the rate increases. For example, at a distance of 10 Mpc the estimated supernova rate is one per year. However, the corresponding amplitude would be $h \approx 3 \times 10^{-21}$ to 3×10^{-20}.

By combining the sensitivity estimates for gravitational wave detectors now under construction and the expected amplitudes and frequencies of possible sources we see that the possibility for detection in the near future is good. Furthermore, as the technology improves, an era of gravitational wave astronomy may soon be possible. Since gravitational waves are not absorbed by intervening matter as is electromagnetic radiation, such an astronomy may allow us to explore regions of the universe, such as the centers of galaxies, that are now blocked to our view.

E. Gravitational Lenses

In many of its effects, a gravitational field acts like a medium with a variable index of refraction. Thus, two of the observed effects discussed above, the bending of light and the time delay of signals as they pass through a gravitational field, can be understood on this basis. A further consequence of this notion is that there should exist gravitational lenses with properties similar to those of ordinary optical lenses. The most likely candidates for such lenses are galaxies. If placed between us and a distant point source such as a quasar, a galaxy can, in effect, provide more than one path along which light from the source may reach the observer. As a consequence, one would see multiple images of the source. Applied to a galaxy, the bending formula (38) (with $\gamma = 1$) gives typical bending angles of about 1 arcsec.

The first gravitational lens, predicted by Einstein in 1936, was observed in 1979. In 1998 a complete ‘Einstein’ ring was observed and the Californian-Arizona Space Telescope Lens survey lists 43 probable examples of lensing including the most distant galaxy so far observed. In addition, a study of the multiple images produced by the light passing through a galactic cluster has enabled astronomers to calculate the mass distribution in this cluster including the contribution due to ‘dark matter.’

VII. GRAVITY AND QUANTUM MECHANICS

A. Hawking Radiation and Black Hole Thermodynamics

For most of its history, general relativity has stood apart from quantum mechanics. Early attempts to quantize the gravitational field proved to be largely unsuccessful and quantum theory usually neglected the presence of gravitational fields. For most problems one could justify this neglect. The radius of the first Bohr orbit of a hydrogen atom held together by gravitational rather than electrical forces, for example, would be about 5×10^{30} m, which is almost four orders of magnitude larger than the radius of the visible universe! However, one is not justified in ignoring the effects of strong gravitational fields, such as those that occur near the Schwarzschild radius of a black hole, on the behavior of quantum systems since gravity couples universally to all physical systems. When one takes account of the gravitational field in the quantum description of a system, qualitatively new features emerge. One such feature is the phenomenon of Hawking radiation.

Even in the vacuum, where there are no real quanta, pairs of virtual quanta of the various matter fields observed in nature are being continually created and destroyed in equal numbers. Their presence is manifested in such phenomena as the Lamb shift in hydrogen and the Casimir effect. According to Hawking, a black hole can absorb one member of such a virtual pair, leaving its partner to propagate as a real quantum of the field. The energy needed for this process to occur is supplied by the gravitational energy of the black hole. Hawking was able to show that, as a black hole formed, such a flux of real quanta should be produced and that it would be equal to the flux produced by a hot body of temperature T given by

$$kT = \hbar g/2\pi c, \tag{51}$$

where k is Boltzmann’s constant, $\hbar$ is Planck’s constant, and g is the gravitational acceleration at the Schwarzschild radius of the black hole and is equal to $c^4/4GM$, where M is its mass.

For a solar mass black hole this temperature would be 2.5×10^{-6} K. However, for a 10^{12} kg mass black hole it would be 5×10^{12} K. Such a black hole would emit energy at a rate of about 6000 MW, mainly in gamma rays, neutrinos, and electron–positron pairs. Hawking has suggested that “primordal black holes” with such masses might have been formed by the collapse of inhomogeneities in the very early stages of the universe and that some of them might have survived to the present day. If so, they probably represent our only hope of observing Hawking radiation. However, measurements of the cosmic-ray background around 100 MeV place an upper limit for black holes with masses around 10^{15} kg of about 200 per cubic light-year.

The fact that a black hole can radiate real quanta might seem to contradict the fact that no radiation can escape from a black hole. However, that restriction is only true classically. One can think of the emitted radiation as having come from inside the event horizon surrounding the black hole by quantum mechanically tunneling through the potential barrier created by its gravitational field. Actually, it is possible for a black hole to emit almost any configuration of quanta, including macroscopic objects. Since we cannot have direct knowledge of the interior of a black hole, all we can determine are the probabilities for the emission of such configurations. The overwhelming probability is that the emitted radiation is thermal with a temperature given by Eq. (50).

That a black hole should have associated with it a temperature fits in with some analogies between black holes and thermodynamics discovered by Bardeen, Carter, Hawking, and Bekenstein. If the energy density of the matter that went to make up the black hole is nonnegative, it can be shown that, classically, the surface area of the event horizon surrounding it can never decrease with time. Moreover, if two black holes coalesce to form a single black hole, the area of its event horizon is greater than the sum of the areas of the event horizons surrounding the two original black holes. These properties are very similar to those of ordinary entropy. Furthermore, when a black hole forms, all information concerning its structure except its mass, charge, and angular momentum is lost, and when ordered energy is absorbed by a black hole it too is forever lost to the outside world. These considerations led Bekenstein to associate with a black hole an entropy S given by

$$S = ckA/4\hbar, \tag{52}$$

where A is the surface area of the event horizon surrounding the hole.

The existence of Hawking radiation raises the question of the ultimate fate of a black hole. As it radiates, its mass decreases. Its temperature therefore increases, and hence so does the rate of emission of radiation. What is left is, at this point, speculation. It might disappear completely, it might cease radiating when its mass reaches some critical value, or it might continue radiating indefinitely, creating a negative-mass naked singularity. While the latter two possibilities seem unlikely, the first one implies that whatever matter went into making the black hole initially would simply cease to exist. In deriving the emission from black holes, the gravitational field was treated classically while the matter fields were treated quantum mechanically. It has been suggested that when the mass of the black hole becomes comparable to the Planck mass (M_P), that is, the mass one can form from the constants $c, \hbar$, and G, namely, $(\hbar c/G)^{1/2} \sim 5 \times 10^{-15}$ g, the gravitational field can no longer be treated as a classical field. If so, the fate of a black hole will be decided only when we have a consistent theory of quantum gravity.

B. Quantum Gravity

The search for a consistent quantum theory of gravity still stands as a major challenge for physics. Because of the extreme weakness of the gravitational force compared to the other fundamental forces in nature, the electro-weak and the strong, it is clear that the quantum effects of gravity would manifest themselves only at extremely short distances and correspondingly high energies. It has been argued that one measure of such a distance is the so-called Plank length L_P formed from G, h, and c and given by

$$L_P = \left(\frac{Gh}{c^3}\right)^{\frac{1}{2}} \simeq 10^{-35}\,\mathrm{m}, \tag{53}$$

since this is the radius at which the Compton wavelength of a black hole is equal to its Schwarzschild radius. To probe such a small length would, it is further argued, require energies of the order of a Plank mass which, in energy units is about 10^{18} GeV. Thus one might expect that the effects of "quantum gravity" would become important for instance when the radius of the early universe was of the order of a Plank length. Put another way, one would expect that the laws of classical gravity would no longer be applicable within such small distances so that any attempt to describe the evolution of the universe during this "Plank" era would require some kind of quantum theory of gravity. Furthermore the search for a "unified" theory of elementary particles could not be considered complete without the inclusion of the gravitational field and since such a theory must, of necessity be a quantum theory, it must include a quantum theory of gravity. The other argument for quantizing the gravitational field is that, if it were strictly a classical field, one would be able to determine both the position and momentum of its sources simultaneously and thus violate the uncertainty principle.

One possible approach to the construction of a quantum theory of gravity is to proceed along the lines used to quantize other fields such as the electromagnetic field. One constructs a Hamiltonian version of the theory and then applies the rules of quantum mechanics. The fields and their conjugate momenta are taken to be operators and one writes down an appropriate Schrödinger equation. However, when one attempts to apply this prescription one runs into apparently insurmountable difficulties. The classical Hamiltonian theory possesses a number of so-called constraint equations, algebraic relations between the field variables and their conjugate momenta, which have been studied by Peter Bergmann and his students and by Paul Dirac. These constraints are generators of the infinitesimal coordinate transformations under which the theory is invariant and as such must satisfy a specific Poisson bracket algebra. Unfortunately, in the quantized version of the theory using the gravitational field as the basic field variables no factor ordering of the constraints satisfies an analogous commutation algebra. Furthermore, only "observables," i.e., quantities that are themselves invariant under the invariant transformations of the theory are quantized. The problem of finding such observables in general relativity is a still an unsolved problem.

If, on the other hand, one proceeds to quantize the linear Einstein equations and treat the nonlinear terms as perturbations one also runs into serious difficulties. As in all quantum field theories one encounters divergent integrals. In successful theories such as quantum electrodynamics where there only a finite number of such integrals, they can be "renormalized" away. However, in the gravitational field case one encounters new divergent integrals at each new order of approximation making the theory non-renormalizable.

In recent times there have been several new attempts to construct a quantum theory of gravity. One such approach is supergravity.

It is an extension of supersymmetric quantum field theory in which every boson in the theory has associated with it a fermion and vice versa. Furthermore, the theory is invariant under the interchange of bosons and fermions. Such theories have been used to construct a unified theory of the strong and electroweak interactions between elementary particles. In supergravity, one associates a spin $3/2$ particle called the gravitino with the quantum of the gravitational field. This theory has been shown to have no logarithmic divergences in the lowest two orders of perturbation theory even when gravity is coupled to matter. Whether supergravity or one of its extensions proves to be a viable theory is a matter for future investigation.

The two most recent attempts to construct a quantum theory of gravity are Superstring theory and canonical (Hamiltonian) quantization using so-called Ashtekar variables. Both approaches show considerable promise at this time. Superstring theory replaces the point particles of quantum field theory by strings, membranes or higher dimensional extended structures called p-branes. In these theories p-brane excitations correspond to particle states and since one of these excitations corresponds to a massless spin 2 state, that state was identified with a graviton, the putative quanta of the gravitational field. Also, the metric used to construct the string or p-brane action must satisfy the Einstein equations, albeit in eleven or more dimensions, in order to insure the cancellation of a dilational anomaly in the theory. Most recently it was shown by A. Strominger and C. Vafa that by counting the energy levels of the superstring gravitational field for a black hole one obtains the Bekenstein formula (52) for the entropy of a black hole. What is still lacking in the theory at this time however is a reduction, in some classical limit, of the full Einstein field equations. Such a derivation would show just how the classical gravitational field arises as an 'emerging' property of a more basic underlying structure in much the same way as the hydrodynamic variables of fluid mechanics arise from the Boltzmann equation, which itself arises from classical many-body theory. It might even show how the very notion of a space-time point arises from such a structure.

The Ashtakar approach to constructing a theory of quantum gravity is much less radical than that taken by string theory. Ashtakar starts from the classical Einstein equations but rather than working with the gravitational field as the fundamental variable he tries to recast the theory into a form resembling a Yang-Mills gauge quantum field theory. In effect he uses the connection which, in terms of the field $g_{\mu\nu}$ is given by Eq. (18) and an additional tetrad field. In addition the connection is complexified, that is, treated as a complex field. They resulting field equations, though equivalent to the Einstein equations as far as their physical content is concerned, have the form of Yang-Mills equations. As in the hamiltonian version of the Einstein equations, the Ashtakar variables satisfy a number of algebraic constraints. These latter constraints however are not as formidable as those between the $g_{\mu\nu}$ and their conjugate momenta and it appears that they can by solved nonperturbatively.

The two approaches, string theory and canonical quantization using Ashtakar variables, appear to be incompatible since the former is based on elementary excitations (gravitons) while the latter, being nonperturbative, does not. At present, however, the physical consequences of neither theory have been explored sufficiently to allow one to decide if indeed they are equivalent or whether in fact either one agrees with observation.

SEE ALSO THE FOLLOWING ARTICLES

Celestial Mechanics • Cosmology • Gravitational Wave Physics • Manifold Geometry • Mechanics, Classical • Mössbauer Spectroscopy

● QUANTUM MECHANICS ● QUANTUM THEORY ● RELATIVITY, SPECIAL ● STELLAR STRUCTURE AND EVOLUTION ● X-RAY ASTRONOMY

BIBLIOGRAPHY

Anderson, J. L. (1967). "Principles of Relativity Physics," Academic Press, New York.

Bergmann, P. G. (1968). "The Riddle of Gravitation," Scribner's, New York.

Bertotti, B., de Felice, F., and Pascolini, A. (eds.) (1984). "General Relativity and Gravitation," Reidel, Dordrecht, Netherlands.

Blair, D. G., and Buckingham, M. J. (eds.) (1989). "Fifth Marcel Grossman Meeting on Recent Developments in General Relativity," World Scientific, New Jersey.

Cooperstock, F. I. (ed.) (1990). "Developments in General Relativity," I.O.P., Bristol, U.K.

Evans, C. R., Finn, L. S., and Hobill, D. W. (eds.) (1989). "Frontiers in Numerical Relativity," Cambridge University Press, Cambridge, U.K.

Hawking, S. W. (1988). "A Brief History of Time," Bantam Books, New York.

Hawking, S. W., and Israel, W. (eds.) (1979). "General Relativity," Cambridge University Press, Cambridge, U.K.

Hawking, S. W., and Israel, W. (eds.) (1987). "300 Years of Gravitation," Cambridge University Press, Cambridge, U.K.

Kaufman, W. J., III. (1977). "The Cosmic Frontiers of General Relativity," Little, Brown, Boston, MA.

Misner, C. W., Thorne, K. S., and Wheeler, J. A. (1971). "Gravitation," Freeman, San Francisco.

Rindler, W. (1977). "Essential Relativity," 2nd ed., Springer Verlag, New York.

Sexl, R., and Sexl, H. (1979). "White Dwarfs, Black Holes," Academic Press, New York.

Smarr, L. (ed.) (1979). "Sources of Gravitational Radiation," Cambridge University Press, Cambridge, U.K.

Thorne, K. S. (1994). "Black Holes and Time Warps: Einstein's Outrageous Legacy," Norton, New York.

Wald, R. M. (1984). "General Relativity," Univ. of Chicago Press, Chicago.

Weinberg, S. (1972). "Gravitation and Cosmology," Wiley, New York.

Wheeler, J. A. (1989). "A Journey into Gravity and Spacetime," Freeman, New York.

Will, C. (1981). "Theory and Experiment in Gravitational Physics," Cambridge University Press, Cambridge, U.K.

Relativity, Special

Kathleen A. Thompson
Stanford University

GLOSSARY

Antiparticle The partner of a subatomic particle which has the same mass but has electric charge (and certain other quantum numbers) of the opposite sign.

Conserved quantity A quantity which, although it may be different in different inertial reference frames, does not change with time as viewed in any particular inertial frame.

Event A point in space–time, i.e., the mathematical idealization of "something happening" at a particular point in space at a particular moment in time.

Inertial frame A reference frame in which Galileo's law of inertia holds—an object that is at rest and subject to no external forces remains at rest, and an object in motion continues to move at a constant velocity. (An inertial frame is also sometimes called a Lorentz frame.)

Invariant quantity A quantity that is the same in all inertial frames.

Mechanics The science of the laws governing the motion of material objects.

Proper length Length as observed in a frame which is at rest with respect to the length being measured.

Proper time Time (between two events) as measured by a clock carried at constant velocity from one event to the other.

Reference frame A spatial coordinate system, along with synchronized clocks which are at rest in that coordinate system.

World line A curve representing a series of events in space–time (for example, the history of a particle).

SPECIAL RELATIVITY is the branch of physics formulated by Albert Einstein in 1905 that successfully describes the motion of objects, even when they are moving at extremely high speeds with respect to each other. At the beginning of the 20th century many physicists were aware that there were some difficulties in existing physics theories. On the one hand there was a theory of mechanics (the theory of motion of bodies) that had been developed by Galileo, Newton, and others in the 17th century. This theory was well established and had many successes. There was also a newer but also very successful theory of electromagnetism that had been put forth by the Scottish physicist James Clerk Maxwell in 1861. However, these two theories were hard to reconcile with certain experimental observations, and they did not seem to fit together consistently.

One symptom of the problems was the fact that the speed of light in empty space seemed to be always the same, no matter how fast the source of the light and/or the observer were moving. Based on Newtonian mechanics (and on everyday experience with velocities), it was expected that if an observer is traveling in the same direction as a beam of light, the light should appear to him/her to have a slower speed than it does to an observer moving opposite to the direction of the light. Of course, since light moves at extremely high speed, it requires very sensitive experiments to accurately measure its speed and look for a dependence on the motion of the source and/or the observer.

As we shall see, it turns out that Newtonian mechanics is really only an approximate theory, useful when the velocities involved are not too high. Special relativity replaces Newtonian mechanics and is consistent with Maxwell's theory of electromagnetism. Although special relativity is elegant, logical, and one of the cornerstones of modern physics, some of its consequences can seem bizarre and paradoxical at first. This is because our everyday experience and perceptions involve objects that have speeds much less than the speed of light. The speed in light in empty space, denoted by c, is equal to 2.99792458×10^8 m/sec (this number is exact, since the meter is now defined to make it so). At this speed, a beam of light can travel all the way around the earth about seven times in a second.

When velocities are much lower than c, the differences between the predictions of special-relativistic and Newtonian mechanics are very slight. However, in situations where the differences between the two theories are large enough to be measured, special relativity has always turned out to be right. Special relativity is now an essential tool in many scientific and technical fields, including nuclear physics and reactors, astrophysics, acceleration and control of high-energy particles, and our fundamental theories of the elementary particles from which the universe is made.

I. OBSERVERS, REFERENCE FRAMES, AND OTHER PRELIMINARIES

The process of making measurements and observations involves some subtleties in relativity. In everyday life we don't normally worry about the distinction between "when an event happens" and "when we see an event happen" because the speed of light is so high that the time for it to travel to our eyes is undetectably small. When dealing with velocities near the speed of light, as in relativity, we must be more careful. So to begin, we need to discuss observers in the context of relativity.

We define an observer's *reference frame* to be a spatial coordinate system in which the observer is at rest, along with clocks at locations of interest. These clocks are at rest in the spatial coordinate system, and we assume a clock is conveniently located in the vicinity of any event whose time of occurrence we need to measure. An *event* is the mathematical idealization of the concept of something happening at a particular place at a particular time. Obviously an event (e.g., a handclap) has an existence independent of anyone's frame of reference. But to specify an event in a given reference frame, we can give three spatial coordinates to tell *where* it occurs and a time coordinate to tell *when*.

The observer can specify where an object is located by giving its spatial coordinates (x, y, z) in his coordinate system. We will use bold-faced notation to denote such three-dimensional vectors, in particular we use $\mathbf{x}$ as a short-hand for the position vector (x, y, z). Our observer also needs a fourth coordinate so that he can specify the time t when the object is at the location (x, y, z). The observer in a given reference frame may specify the motion of an object by giving its three spatial coordinates (x, y, z) in that frame as a function of the time. We imagine that the time of arrival at a given spatial location is to be read from a nearby ("local") clock.

The observer (or observers) in a given reference frame are assumed to have access to the data in a given frame, even if the events being measured are distantly located. However, if the space and time coordinates of a number of events are recorded in that frame, it may take some time for an observer located at a given place to gather together all that data. The observer cannot instantaneously find out what is happening at distant locations—it takes time to transmit information to him. For instance, if a lightning flash occurs at time t at a distance D from the observer, he will not actually see the lightning flash with his eyes until time $t + D/c$, where c is the speed of light. Nevertheless, we say he observes it to occur at time t, since he is assumed to be able to measure how far it is away, and he can take the light travel time into account. Alternatively, another observer in the same frame could record the time on a local clock in that frame (i.e., from a frame clock located very near the lightning flash). Assuming the clocks have been properly synchronized (we will discuss this in more detail later), he/she can simply report the time read from that clock.

Typically we shall be concerned with reference frames moving at constant velocity with respect to each other. The *velocity* is a vector v representing the rate of change of position with respect to time, i.e.,

$$\mathbf{v} = d\mathbf{x}/dt = (dx/dt, dy/dt, dz/dt). \tag{1}$$

The *speed*, which is just the magnitude of the velocity, is

$$|\mathbf{v}| = \sqrt{(dx/dt)^2 + (dy/dt)^2 + (dz/dt)^2}. \tag{2}$$

The *acceleration* is the rate of change of velocity, i.e.,

$$\mathbf{a} = d\mathbf{v}/dt = d^2\mathbf{x}/dt^2. \tag{3}$$

When an object is moving at constant velocity, its acceleration is therefore zero.

Wherever possible, we will keep things simple by choosing coordinate systems so that all motion takes place along just one axis, which we will take to be the x-axis. Suppose there is a second observer moving along the x-axis at constant velocity v with respect to the first observer. [NOTE: *Here and elsewhere in this article, when we are talking about motion confined to a line, typically chosen to be the x-axis, we will represent the velocity by a nonboldface symbol (in this case v) which, unlike a speed, does have a sign according to whether the motion is in the positive or negative x-direction.*] We may attach a second coordinate system to the second observer. To distinguish space and time coordinates in this system from those in the first system, we put a prime on them, i.e., we write x', y', z', t' instead of x, y, z, t.

Throughout this article we shall often use two such coordinate systems oriented so that corresponding axes are parallel. The “primed” frame is moving with constant velocity v along the x-axis of the unprimed frame (see Fig. 1). We choose our reference of time in each system so that $t = t' = 0$ at the moment when the origins of the two spatial coordinate systems are at the same place.

II. HISTORICAL BACKGROUND AND MOTIVATION

A. Pre-relativity Mechanics of Newton and Galileo

1. The Galilean Transformation

As a simple example, suppose the primed frame is fixed with respect to a passenger coach on a train, and the unprimed frame is fixed with respect to the train tracks (assumed to be straight). We choose the x direction to be along the tracks. If a ball is thrown forward at speed $dx'/dt' = 30$ km/hr by a passenger sitting on the train, and the speed of the train with respect to the tracks is $v = 90$ km/hr, then the ball will have speed $dx/dt = 120$ km/hr as observed by a person standing beside the tracks.

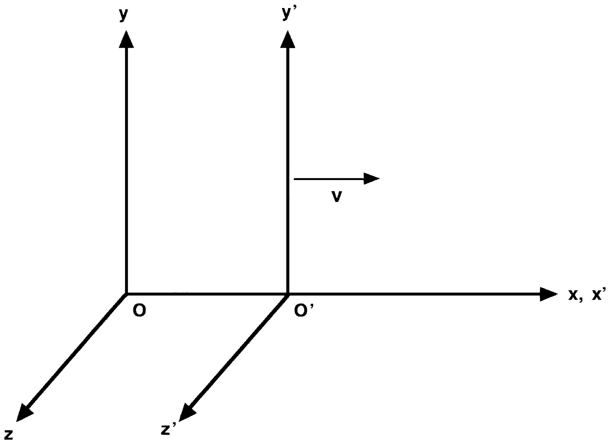


FIGURE 1 Two reference frames in motion with respect to each other at constant velocity. Coordinates are chosen so that the primed frame (x', y', z') moves with velocity v along x-axis with respect to unprimed frame (x, y, z).

This conclusion follows not only from common sense but also from an essential part of Newtonian mechanics—the Galilean transformation. Given the two coordinate systems described above, the Galilean transformation takes the form:

$$\begin{aligned} x' &= x - vt \\ y' &= y \\ z' &= z \\ t' &= t. \end{aligned} \tag{4}$$

This transformation makes explicit some of our common sense ideas about space and time. Time flows at the same rate for observers in both frames. Furthermore, differentiating the first equation with respect to time (and using the fact that, from the last equation, $d/dt = d/dt'$) shows that if the Galilean transformation is valid, then velocities along the same line of motion combine as we expect from everyday experience:

$$dx'/dt' = dx/dt - v. \tag{5}$$

In other words, if an object has velocity $V' = dx'/dt'$ along the x axis according to an observer in the primed frame, then its speed along the x-axis according to an observer in the unprimed frame will be $V = V' + v$.

Before special relativity, it was generally assumed that the Galilean transformation is exactly true. For the train and ball example, it is an extremely good approximation. But suppose the train were instead a very futuristic rocket moving away from some star at 90% of the speed of light. Suppose also that the ball were moving toward the front of the rocket at 30% of the speed of light (with respect to the rocket). Then, as we shall see, it would be completely wrong to conclude that the ball is moving away from the star at 120% of the speed of light.

2. Mass, Energy, and Momentum in Newtonian Mechanics

Before proceeding further into relativity we need to briefly review the concepts of mass, energy, and momentum in Newtonian mechanics. The *mass* is a measure of the inertia of a body, which may be thought of as its "resistance" to the action of a force. In Newtonian mechanics, the total amount of mass in an isolated system (that is, a system that has negligible interaction or exchange with the rest of the universe) is a *conserved* quantity, i.e., it does not change with time. Also, the mass of an object does not depend on how fast it is moving or on its temperature.

In Newtonian mechanics, the *momentum* **p** of an object is defined to be $m\mathbf{v}$ where **v** is the velocity of the object and m is its mass. Thus momentum is a vector with three components

$$p_x = m\,dx/dt, \quad p_y = m\,dy/dt, \quad p_z = m\,dz/dt. \qquad (6)$$

The total amount of momentum in an isolated physical system is also conserved. However, the amount of momentum depends on the reference frame. As a very simple example, consider a single object with mass m that is at rest in some reference frame. If it is isolated from all outside influences, it remains at rest—its momentum is zero. In a reference frame moving with velocity **v** with respect to the object's rest frame, the object has momentum equal to $-m\mathbf{v}$, also unchanging with time.

A system consisting of several objects interacting with each other can have the objects' momenta redistributed, but the sum of the momenta remains constant. For example, in a system consisting of two bodies, the Newtonian law of momentum conservation says

$$m_1\mathbf{v}_{1,i} + m_2\mathbf{v}_{2,i} = m_1\mathbf{v}_{1,f} + m_2\mathbf{v}_{2,f}, \qquad (7)$$

where the subscript "i" labels quantities before the collision and the subscript "f" labels quantities after the collision.

There are two basic forms of *energy*: (1) kinetic energy, the energy of motion of a body, which is defined in Newtonian mechanics by $\frac{1}{2}mv^2$; and (2) potential energy, which may be regarded as "stored" energy. A common example of potential energy is the energy a ball has by virtue of being held above the ground, in the gravitational field of the Earth. A soon as the ball is let go, it accelerates toward the Earth, gaining kinetic energy and losing an (approximately) equal amount of potential energy. The reason we say "approximately" is that the friction of the air through which the ball falls dissipates a small amount of the energy as heat. Heat is "internal" kinetic energy, i.e., the energy of motion of the molecules comprising the ball and the surrounding air. In an isolated system, the total sum of all the different forms of energy is a constant in any given frame of reference. This is the law of conservation of energy.

Newton's Second Law tells how the momentum **p** of an object changes with time when a force **F** is applied:

$$\mathbf{F} = d\mathbf{p}/dt. \qquad (8)$$

In Newtonian mechanics, this can also be written $\mathbf{F} = m\mathbf{a}$, where $\mathbf{a} = d\mathbf{v}/dt$ is the acceleration of the object.

3. Galilean Relativity Principle

A principle of relativity had been formulated by Galileo long before Einstein developed the special theory of relativity. Galileo considered the example of a ship in smooth waters, carrying passengers who cannot see out, but who have with them a bowl of fish, a bottle dripping water into a bowl, and some birds and butterflies. From their observations, the passengers cannot tell whether the ship is at rest or moving forward in a straight line at constant speed. As Galileo said,

> . . . the little animals fly with equal speed to all sides of the cabin. The fish swim indifferently in all directions, the drops fall into the vessel beneath, and in throwing something to your friend you need throw it no more strongly in one direction than another. . . .

More formally, this means that the if the fundamental laws of motion hold in one reference frame, then they also hold in any reference frame moving at constant velocity with respect to the first frame. This is the *Galilean relativity principle*.

To see one example of this, consider the law of momentum conservation as expressed in Eq. (7). If we change to a frame of reference moving with velocity **V** with respect to the original frame, then according to the Galilean transformation we simply subtract **V** from each of the velocities in Eq. (7). to get the velocities in the new frame ($\mathbf{v}'_{1,i} = \mathbf{v}_{1,i} - \mathbf{V}$, etc.). The masses m_1 and m_2 are not changed by transforming to a different reference frame. Thus, transforming to the new frame is equivalent to subtracting $(m_1 + m_2)\mathbf{V}$ from both sides of Eq. (5). So, momentum conservation also holds in the new frame:

$$m_1\mathbf{v}'_{1,i} + m_2\mathbf{v}'_{2,i} = m_1\mathbf{v}'_{1,f} + m_2\mathbf{v}'_{2,f}. \qquad (9)$$

As another example, we can check that Newton's Second Law is invariant under Galilean transformations. Differentiating Eq. (5) gives $d^2x'/dt'^2 = d^2x/dt^2$, since $t = t'$ and v is constant. This just says that $a = a'$, i.e., the acceleration of an object is the same as measured in any two frames moving at constant velocity with respect to each other. Furthermore, in Newtonian mechanics, the masses and forces are the same in the two frames. Thus Newton's Second Law is valid in the primed frame, if it is valid in the unprimed frame.

B. Propagation of Light

Maxwell's theory of electromagnetism showed how electric and magnetic fields are related to each other and to the presence and motion of electric charges. The heart of the theory is Maxwell's equations for the fields, along with an equation (Lorentz electromagnetic force law) that gives the force on a charge in terms of the local electric and magnetic field and the velocity of the charge. In Maxwell's theory, light is a wave consisting of undulating patterns of electric and magnetic fields. Visible light is just a small slice of a spectrum of electromagnetic waves ranging from long wavelength radio waves to very short wavelength gamma rays. We use the term "light" to refer to electromagnetic waves of any wavelength, even if they are outside the range visible to our eyes.

Special relativity was originally motivated by the observation of several fairly subtle effects involving light. To understand the issues involved, we must review some of what was known about light. Early in the 19th century (1801–1804), Thomas Young carried out a quantitative demonstration of interference in light, using a double-slit experiment. Interference is characteristic of waves—it simply means that when two waves interact, they reinforce or cancel each other depending on their relative phases of oscillation. Young's experiments were followed by detailed studies of interference, diffraction, and polarization of light, by A. J. Fresnel and others. Thus by Maxwell's time, it was well established that light exhibited many of the properties of a wave.

1. The "Luminiferous Ether" Hypothesis

All waves familiar at that time (e.g., sound waves, water waves) were known to take place in a material medium. All these waves have the property that their apparent speed depends on the motion of the observer with respect to the medium. It was natural to assume that there exists some kind of medium to "carry" light waves. It was difficult to account for the properties of such a medium in terms of a mechanical model. It was known that the speed of light was very high (which would imply that the medium exerts very strong restoring forces when displaced from equilibrium and yet there was no direct evidence for its existence). Nevertheless, it was assumed by many people that there does exist a "luminiferous ether" permeating all space, to play this role. It was expected that the speed of light should appear different depending on the observer's motion with respect to the ether.

The ether hypothesis ran into difficulties, although it managed to survive a number of experimental tests by introducing further hypotheses. For example, in order to avoid contradictions with some of the experiments, it was necessary to assume that the ether is partially dragged along with moving material media in a very specific way that depends on the index of refraction of the medium.

2. The Michelson–Morley Experiment

It was important to look for direct evidence of the preferred frame of reference that the ether was supposed to provide (if it existed at all). The most famous such attempts were carried out by A. A. Michelson and E. W. Morley. Their goal was to detect an effect on the measured speed of light due to motion of the observer through the ether. The experiments used an interferometer invented by Michelson (1881), and later refined in collaboration with Morley (1887).

The basic design of the interferometer is shown in Fig. 2. It is essentially a device in which a beam of monochromatic (single-wavelength) light is split, follows two

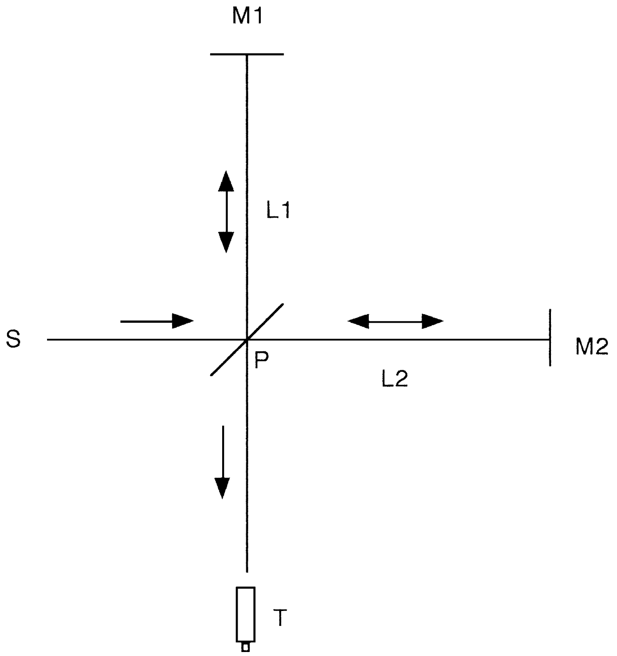


FIGURE 2 Schematic of Michelson–Morley experiment.

different paths, and is recombined. A small difference in the two optical pathlengths (i.e., a difference in the number of wavelengths of light along the two paths) can be detected by looking at the interference pattern created by the recombined beams. The beam of monochromatic light originates from a source S. The plate P partly reflects and partly transmits the light. Thus there is one path that reflects from P to mirror M_1 and goes back through P to the observation telescope T. There is a second path that goes through P to the mirror M_2, is reflected back to P where it is reflected to the observation telescope T. The distances from P to M_1 and to M_2 respectively are L_1 and L_2.

When the interferometer is rotated the optical path difference along the two arms would change, if the ether hypothesis is correct. The expected change depends on its speed v with respect to the ether and on its orientation with respect to the direction of motion. For example, if the interferometer is first oriented so that one arm is pointing along the direction of motion through the ether, and is then rotated by 90° so that the other arm is along the direction of motion, one can show that the change in optical path difference would be

$$\Delta\delta \approx \frac{(L_1 + L_2)v^2}{\lambda c^2}, \tag{10}$$

where λ is the wavelength of the light. Michelson and Morley expected that for at least some orientations of the apparatus and some times of year, the speed v of the apparatus through the ether should be greater than or equal to the earth's orbital speed of about 30 km/sec. The resulting value of $\Delta\delta$ should have been large enough for them to detect, but their result was null—no change in the optical path difference was observed, no matter how the apparatus was rotated.

C. Lorentz–Fitzgerald Contraction and Lorentz Transformations

1. Lorentz–Fitzgerald Contraction

It appeared to be impossible to detect and measure motion with respect to an ether. Shortly after the conclusion of the Michelson–Morley experiments, a possible explanation for their null result was proposed separately by both H. A. Lorentz and G. F. Fitzgerald. Their suggestion was that an object is contracted along its direction of motion by the factor $(1 - v^2/c^2)^{1/2}$, where v is the speed with respect to the ether. This would lead to a difference in the lengths of the two arms of the interferometer that would be just right to cancel the expected effects of motion through the ether. At first this proposal seemed to be little more than another *ad hoc* assumption introduced to keep alive the notion of an ether.

2. Lorentz Transformations

Lorentz and other physicists had also been also troubled by the fact that Maxwell's equations are not invariant when a Galilean transformation of the coordinates is applied. However, Lorentz and the mathematician H. Poincare noticed that Maxwell's equations were invariant under a different transformation, whose significance and fundamental nature were not clear at the time. This transformation, which has come to be known as the Lorentz transformation, is as follows:

$$\begin{aligned} x' &= \gamma[x - vt] \\ y' &= y \\ z' &= z \\ t' &= \gamma[t - (v/c^2)x]. \end{aligned} \tag{11}$$

Here γ is defined by:

$$\gamma \equiv \frac{1}{\sqrt{1 - v^2/c^2}}, \tag{12}$$

and is called the *Lorentz factor*. Note that $\gamma \geq 1$ and that $1/\gamma$ is the Lorentz–Fitzgerald contraction factor. As discussed earlier for the Galilean transformation, we have set up our coordinate systems so that the primed system with respect to the unprimed system has velocity v along the x axis, and the origins of the two coordinate systems coincide at time $t = 0$. At the time, this invariance of Maxwell's equations under the Lorentz transformation was intriguing, but the full implications of the transformation were not understood. It does predict an apparent contraction along the direction of motion, just as is needed to explain the null result of the Michelson–Morley experiment. However, if the last of Eq. (11) is correct, time flows differently for different observers—a major departure from Newtonian mechanics!

The Lorentz transformation, if valid, also suggests that the constant c, the speed of light in vacuum, is a "universal speed limit." As v approaches c, the factor γ approaches infinity. If v were to become larger than c, γ would be the square root of a negative number, which suggests that $v > c$ does not occur in reality. All existing experiments and observations are indeed consistent with the hypothesis that no material object or causal influence propagates faster than c.

One of the key contributions of Einstein was to derive the Lorentz transformation from basic principles and to show that it, rather than the Galilean transformation, is the correct way to relate the coordinates in two reference frames in uniform motion with respect to each other. He then proceeded to build a theory that ended up revising many of the accepted ideas in physics.

3. Invariance of Transverse Distances

As we have just discussed, physicists were forced to consider the possibility that observed lengths may be contracted along the direction of motion. Note that in both the Galilean and the Lorentz transformations, distances transverse (i.e., at right angles) to the direction of motion do not change. The necessity for this invariance of transverse dimensions follows from simple symmetry considerations.

For example, suppose a piston fits exactly inside a cylinder when it is at rest. Then suppose the piston and cylinder are moving toward each other at very high speeds as shown in Fig. 3. What happens when the piston and cylinder meet? We might choose to analyze the situation in either the rest frame of the piston or the rest frame of the cylinder. If the observed transverse coordinates of a moving object either grew or shrank, then according to one of the frames the piston would be smaller transversely, and should be able to sail into the cylinder and make a dent in the back wall. In the other frame, the piston would have the larger diameter, so it would collide with the outer rim of the cylinder and never reach the back wall at all. But the result cannot depend on the observer's frame—either there is a dent on the back wall or there isn't! So the assumption that the transverse dimensions depend on the frame must be wrong.

Even more fundamentally, there is no sensible way to define an axis toward which the transverse coordinates in a given frame would grow or shrink—it was really quite arbitrary to choose this to be the axis of symmetry of the piston and cylinder in Fig. 3. In summary, if we assume the direction of relative motion is along the x-axis, there is no consistent way to define a transformation of y and z other than the identity transformation

$$\begin{aligned} y' &= y \\ z' &= z. \end{aligned} \tag{13}$$

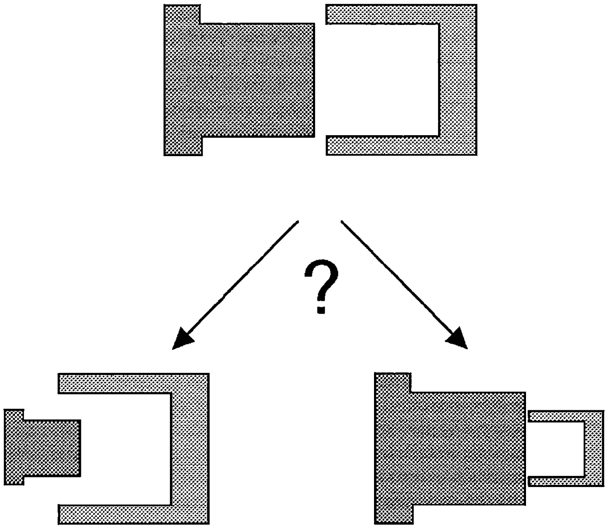


FIGURE 3 A situation illustrating the invariance of transverse dimensions.

III. FOUNDATIONS OF SPECIAL RELATIVITY

A. Inertial Reference Frames

In order to develop relativity, we first need to revisit the concept of the observer and his reference frame. Specifically, we need to define what we mean by an "inertial reference frame."

Galileo postulated, based upon simple experiments, that a body which is moving at constant velocity (i.e., in a straight line at constant speed) continues to move with that same velocity provided that no forces act upon it. This is known as *Galileo's law of inertia* (and was later postulated by Newton as his First Law). We define an *inertial reference frame* to be a reference frame in which Galileo's law of inertia holds.

One example of an inertial frame is an unpowered space capsule drifting freely through space. The interior of an elevator that has broken loose from its cable and is freely falling in the gravitational field of the earth is also a good approximation to an inertial frame until the elevator crashes into the ground. An observer inside the falling elevator will find that if an object is given a small push, it will move in a straight line as seen in a spatial coordinate system fixed in the elevator.

Strictly speaking, an inertial frame is always only an approximation to reality since it can never be completely free of extraneous influences. If the observer's measurements are sensitive enough, deviations from Galileo's law of inertia can in general be detected. In the freely falling elevator, if position and time measurements are sensitive enough it will be possible to observe small effects due to the fact that the strength and direction of the gravitational field are very slightly different at different locations in the elevator.

In the context of both Newtonian mechanics and special relativity, it is common to think of an inertial frame as a frame that is moving with "constant velocity." Of course (as Newton himself realized) this leaves open the question "constant velocity with respect to what?" With the above definition of an inertial frame, we are able to leave aside the question of whether motion at constant velocity has a meaning in any absolute sense (for example, with respect to the average distribution of matter in the universe). There is no requirement that the motion be at constant velocity with respect to the rest of the universe. The freely falling elevator is an example of an inertial frame that is not moving at constant velocity—it is accelerating toward the center of the earth. All that we require of an inertial reference frame is that Galileo's law of inertia hold to within the sensitivity of our measurements, in the region of space and for the duration of time that we are concerned with.

Hereafter when we say "reference frame" or just "frame," we shall mean an inertial reference frame unless otherwise specified.

B. Postulates of Relativity

According to the Galilean relativity principle, the laws of motion are the same in all inertial frames. Einstein extended the relativity principle to ALL the laws of physics, not just the laws of motion, and took it as one of the basic postulates of his special theory of relativity.

1. Postulate 1 (Principle of Relativity): The fundamental laws of physics are the same in every inertial reference frame.

Einstein, in recalling a paradox he had first thought about when he was sixteen, wrote:

> If I pursue a beam of light with the velocity c (velocity of light in a vacuum). I should observe such a beam of light as a spatially oscillatory electromagnetic field at rest. However, there seems to be no such thing, whether on the basis of experience or according to Maxwell's equations.

In other words, Einstein began to question whether it was possible even in principle to "catch up with" a beam of light. This, along with the failure of all experiments to detect any changes of the speed of light in empty space, motivated him to assume.

2. Postulate 2 (Invariance of the Speed of Light): The speed of light in vacuum is the same in all reference frames.

In particular, the observed speed of light depends neither on the speed of the source of light nor on the speed of the observer. As noted earlier, Newton's Laws were assumed to be valid in all inertial frames, and these laws of motion are invariant under Galilean transformations. Special relativity modifies this by assuming that ALL correct fundamental laws of physics are valid in all inertial frames, and the fundamental laws of physics are invariant under Lorentz transformations. Newton's Laws and the Galilean transformation are good approximations in many situations, but fail badly when dealing with speeds close to the speed of light.

C. Synchronization of Clocks

In order to make meaningful comparisons of time at different locations in a given frame, it is necessary to synchronize the clocks being used in that frame. Since the speed of light c is constant in any frame if Postulate 2 is true, we could proceed as follows. Choose one of the clocks to be the reference clock, and set its time to zero. Set the time on each of the other clocks to time D_i/c, where D_i is the distance from the reference clock to clock i, and hold clock i's time at this value. Now let a flash of light be emitted from the reference clock, and at the same time let it begin to run starting at time $t = 0$. At the moment when the flash arrives at any other clock, let it begin to run starting at its preset value. In this way, all the clocks will be synchronized, since the time lag needed for the flash to reach each clock will be exactly the preset value for that clock. It is important to note that this procedure synchronizes the clocks *according to observers at rest in a particular reference frame*. As we shall see, the clocks will be noticeably out of synchronization according to observers in a frame moving at high speed with respect to the first frame.

D. Time Dilation in Relativity

The postulate that the speed of light in empty space is constant is the basis for many of the results in relativity that go against our normal intuition. As one example of this, consider the following situation. Suppose that there are two parallel mirrors a distance D apart. We can use them to make a clock consisting of a light flash that bounces back and forth between the two mirrors. The clock "ticks" each time the light flash hits either mirror. Suppose that this clock is put on a high-speed rocket.

First consider how the clock looks to an observer (whom we shall call O′) who is moving with the rocket and is at rest with respect to the clock. Observer O′ sees the light moving straight up and down along the distance D between mirrors as shown in Fig. 4(a). According to observer O′, each tick of the clock takes a time $T' = D/c$.

Let the rocket be moving to the right at speed v with respect to another observer whom we shall call observer O (see Fig. 4(b)). Common sense would lead us to expect that each tick of the clock takes the same time for observer O as it does for O′, but from the following arguments we see that this cannot be true if relativity is valid.

First, the dimensions transverse to the direction of motion are invariant. In the present situation this means that observer O agrees with observer O′ that the distance between the two mirrors is D.

Next, we note that while the light is going from one mirror to the other, the rocket moves to the right in observer

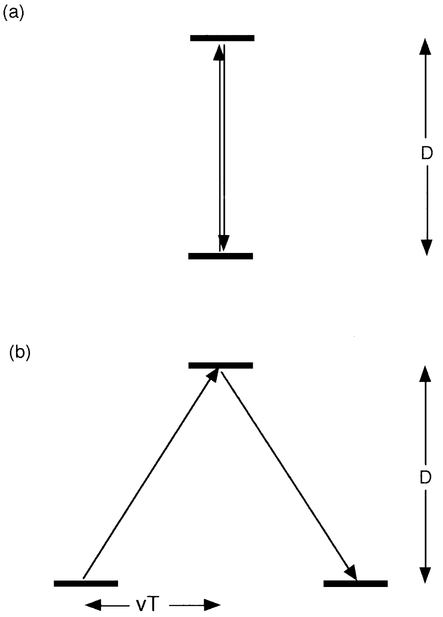


FIGURE 4 Clock consisting of a light flash bouncing back and forth between two mirrors. (a) Path of light flash as seen by an observer at rest with respect to the clock. (b) Path of light flash as seen by an observer with respect to whom the clock is moving to the right with speed *v*.

O's frame by the distance vT, where v is the speed of the rocket with respect to observer O and T is the time that observer O says it takes the light to travel between mirrors. So according to observer O, the light must travel a longer path between mirrors than it travels according to observer O′. From Einstein's second postulate, the speed of light has the same value c in the frame of the observer O as it does for the observer O′. Therefore, observer O will say that a tick of the clock takes longer than does observer O′!

It is straightforward to derive the factor by which the time per clock tick differs for these two observers. From the Pythagorean theorem, the distance travelled by the light during one tick, as seen by observer O, is $\sqrt{D^2+v^2T^2}$. Thus the time for one tick is $T=\sqrt{D^2+v^2T^2}/c$. Using $D=cT'$ to eliminate D, we find

$$T=\frac{1}{\sqrt{1-v^2/c^2}}T'=\gamma T', \tag{14}$$

where γ is the Lorentz factor. The faster the rocket goes, the more slowly the clock ticks, according to observer O.

Although we have been talking about a particular type of clock, it follows from the principle of relativity that all other clocks must run slow by the same factor. For if they did not, then the very fact that clocks get out of synchronization in a moving frame could be used to distinguish a moving frame from one at rest. This would contradict the postulate that the laws of physics are the same in all inertial frames.

Furthermore, there is no inherent asymmetry between observers O and O′. From the point of view of O′, it is O who is moving (at velocity $-v$ along the x-axis). By the same kind of argument as given above, observer O′ will observe clocks that are moving with O to run slow (by the factor γ) compared to his own clocks.

E. Derivation of the Lorentz Transformation

The Lorentz transformation, originally postulated in an *ad hoc* manner to explain the Michelson–Morley experiment, can now be derived. Assuming Einstein's two postulates, we now show that the Lorentz transformation is the only possible transformation between two inertial coordinate systems moving with constant velocity with respect to each other.

The transformation must be linear in the time and space coordinates because of the Principle of Relativity (Postulate 1). If the transformation were not linear, then uniform motion in a straight line in one frame would no longer appear as uniform straight-line motion in another frame moving at constant velocity with respect to the first. This would contradict the requirement that Galileo's law of inertia hold in all inertial frames.

We use the fact that the transformation must not change the coordinates transverse to the axis of relative motion of the two frames (which, as usual, we take to be along the x and x' axes for simplicity). Then $y'=y$ and $z'=z$, just as in the Galilean transformation. With this choice, the transformation equations for x and t must be independent of the transverse coordinates by symmetry (there is no way to single out a particular location or direction of rotation relative to the axis of motion).

Therefore the transformation in x must be of the form

$$x=Ax'+Bt'. \tag{15}$$

Our analysis of the light-flash clock showed that when $x'=0$ we have $t=\gamma t'$ so that $x=v\gamma t'$. This is consistent with Eq. (15) only if $B=\gamma v$. Furthermore, the motion of the origin of the unprimed system ($x=0$), as expressed in the coordinates of the primed system, is given by $x'+vt'=0$. Again comparing with Eq. (15) we must have $B/A=v$, resulting in

$$x=\gamma(x'+vt'). \tag{16}$$

Since we can invert the roles of the primed and unprimed coordinates by reversing the sign of v, we must also have

$$x'=\gamma(x-vt). \tag{17}$$

Equations (16) and (17) may be solved for t in terms of the primed variables:

$$t = \gamma\left(t' + \frac{v}{c^2}x'\right) \tag{18}$$

and for t' in terms of the unprimed variables:

$$t' = \gamma\left(t - \frac{v}{c^2}x\right). \tag{19}$$

Thus we have reproduced the Lorentz transformation given previously as Eq. (11). It may be summarized in a slightly different form—it is often useful to regard all four spacetime dimensions as having the same units, either conventional length units (e.g., meters) or conventional time units (e.g., seconds). We can do this by multiplying the time in conventional units by c. Then ct would be replaced by t (and ct' by t') in Eq. (11). Or we could equally well divide each of the space dimensions in conventional units by c. In either case, the Lorentz transformation with time and space expressed in the same units takes the simple form

$$\begin{aligned} x' &= \gamma(x - \beta t) \\ y' &= y \\ z' &= z \\ t' &= \gamma(t - \beta x). \end{aligned} \tag{20}$$

Here $\beta \equiv v/c$ is the velocity of the primed frame with respect to the unprimed frame, expressed as a fraction of the speed of light, and $\gamma = 1/\sqrt{1-\beta^2} = 1/\sqrt{1-v^2/c^2}$ is the Lorentz factor. The inverse transformation, again with time and space in the same units, is obtained by simply reversing the sign of v (and thus of β) when the roles of the primed and unprimed cooordinates are reversed:

$$\begin{aligned} x &= \gamma(x' + \beta t') \\ y &= y' \\ z &= z' \\ t &= \gamma(t' + \beta x'). \end{aligned} \tag{21}$$

Note that the Lorentz transformation reduces to the Galilean transformation when $v \ll c$ and $x/t \ll c$.

F. The Invariant Interval

The space and time coordinates differ in different frames, but there is an important function of the coordinates that is an invariant, i.e., the same in all frames. This quantity is called the *space–time interval* (or just *interval*) between two events. For two events with space–time coordinates (t_1, x_1, y_1, z_1) and (t_2, x_2, y_2, z_2) we define the square of the interval by

$$\begin{aligned} (\Delta s)^2 &= c^2(t_1 - t_2)^2 - (x_1 - x_2)^2 \\ &\quad - (y_1 - y_2)^2 - (z_1 - z_2)^2 \\ &\equiv (c\Delta t)^2 - (\Delta x)^2 - (\Delta y)^2 - (\Delta z)^2. \end{aligned} \tag{22}$$

With a bit of algebra one can show that the interval is invariant under Lorentz transformations. In other words, the interval between two events is the same regardless of which inertial frame it is calculated in:

$$\begin{aligned} &(c\Delta t)^2 - (\Delta x)^2 - (\Delta y)^2 - (\Delta z)^2 \\ &\quad = (c\Delta t')^2 - (\Delta x')^2 - (\Delta y')^2 - (\Delta z')^2 \end{aligned} \tag{23}$$

if the primed and unprimed coordinates are related by a Lorentz transformation.

If $(\Delta s)^2 > 0$, we say the interval is *time-like*. When the interval between two events is time-like it is possible for an observer to be present at both events, since he does not need to travel faster than the speed of light c to get from one event to the other. If $(\Delta s)^2 = 0$, we say the interval is *light-like*—a light ray can depart from Event 1 and arrive exactly at the right time to be present at Event 2, or vice versa. If $(\Delta s)^2 < 0$, we say the interval is *space-like*. When the interval between two events is space-like then no object or signal can get from one event to the other because it would be necessary to exceed the speed of light. Since the speed of light is assumed to be the maximum speed with which any physical influence can propagate, there cannot be a causal connection between two such events. Furthermore, it can be shown that if two events have spacelike separation, then and only then is it possible for observers in different reference frames to disagree about the time ordering of the two events.

G. Minkowski Diagrams and World Lines

Diagrams in space–time are often referred to as Minkowski diagrams, after H. Minkowski who pointed out a "geometric" way of looking at relativity. The concept of distance in ordinary space is replaced by the concept of the interval between two events in space–time. Just as distance between two points in ordinary space is an invariant regardless of rotations of the coordinate system, the interval between two events in Minkowski space (i.e., space–time) is an invariant regardless of the inertial frame. Of course, the presence of the relative minus sign between space and time coordinates in the definition of interval means that the geometry of Minkowski space is fundamentally different from that of ordinary Euclidean space.

Figure 5 is an example of such a diagram (where for simplicity only one of the three space dimensions is depicted. The entire "history" of a moving particle is represented by

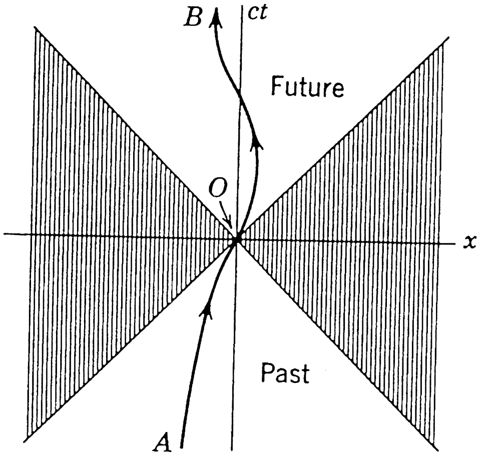


FIGURE 5 A Minkowski diagram (including only one spatial dimension), showing the world-line of a particle and the past and future light-cones of the event at O. [Reproduced with permission from Jackson, J. D. (1975). "Classical Electrodynamics," 2nd ed., p. 519, Wiley, New York.]

a curve in space–time. This curve is called the particle's *world line*. An example of a world line starting at event A and proceeding to event B is shown in Fig. 5.

Note that the absolute value of the slope of a world line must never be less than one (when plotted with time on the vertical axis as shown), since the particle may not exceed the speed of light. For a particle which passes through O, all points in the white region marked "future" are in principle reachable by the particle at later times, since it could get to any of these events without exceeding the speed of light. Similarly, a particle at O could in principle have taken a path that allowed it to have been present at any event in the white region marked "past." However, all events in the cross-hatched region are inaccessible in the sense that they cannot causally influence, nor be influenced by, an event at O. This region is sometimes referred to as "elsewhere" with respect to event O.

H. Proper Time

We saw earlier that an observer who is in motion with respect to a clock will always observe it to run slower than clocks at rest in his own frame. This leads us to the notion of *proper time*. Suppose that the time between two events is to be measured. Assume also that the space–time interval between the two events is timelike, so that is possible for a clock to be present at both events. If the clock is carried at constant velocity from Event 1 to Event 2, then the time between the events as read by that clock is called the *proper time* between the events, i.e., τ is equal to $t_2 - t_1$ in that frame (we use the Greek letter tau (τ) to represent proper time). Both events occur at the same place in the rest frame of the proper clock, that is, the spatial separation $(\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2$ between the two events is zero. Therefore the proper time $\Delta\tau$ is simply related to the space–time interval between two time-like events by

$$\Delta\tau = \Delta s/c. \tag{24}$$

Carrying a clock at constant velocity from one event to another means that the portion of the clock's world line between the two events is a straight line in Minkowski space. If the Minkowski diagram is plotted in the rest frame of the clock, the line is vertical—there is no change in any of the spatial coordinates anywhere along the path—not only is $\Delta x = \Delta y = \Delta z = 0$ for the path as a whole, but $dx = dy = dz = 0$ for any short segment along the path. Thus, in this frame

$$d\tau = dt \tag{25}$$

for each short segment along the path.

Suppose the clock instead takes a less direct route, but is still present at both events, i.e., the clock still departs from x_1, y_1, z_1 at time t_1 and arrives at x_2, y_2, z_2 at time t_2, but it does not travel at constant velocity. In this case it will not be true that $dx = dy = dz = 0$ for all segments along the path. The total elapsed time on the clock is obtained by integrating the proper time along the new path. We assume that the proper time along each short space–time segment of the path may be calculated treating each sufficiently small segment of the world-line as a straight line, and we calculate the proper time assuming the clock moves with constant velocity along each short segment. The proper time along each such segment is then

$$d\tau = \sqrt{(cdt)^2 - (dx)^2 - (dy)^2 - (dz)^2}/c. \tag{26}$$

Comparing segments that range over the same values of the time coordinates, it is obvious that the proper time along the indirect path (Eq. (26)) is less than the proper time along the direct path (Eq. (25)) We see that taking a more circuitous route in spacetime will make the integrated proper time *less* than it is if the clock takes the most direct route! This peculiar feature of the geometry of space-time is due to the relative minus sign between the space and time coordinates.

IV. CONSEQUENCES OF THE LORENTZ TRANSFORMATION

A. Proper Length and Lorentz Contraction

A well-known result of relativity is that objects are observed to be shorter along their direction of motion than when they are at rest. This follows directly from the

Lorentz transformation, but may also be derived by the following simple argument involving time dilation.

The *proper length* of an object is defined to be its length as measured in its own rest frame, Suppose the object is a straight stick having proper length L_0. Let a bee fly directly from one end of the stick to the other at constant speed v with respect to the rest frame of the stick. Then according to an observer in the stick rest frame, the trip will take a time $t_0 = L_0/v$. The bee sees the stick going backwards, also at speed v. But according to the bee, from our previous discussion of time dilation, the trip will take less time—the time as measured in the frame of the bee is $t_{bee} = t_0/\gamma = t_0\sqrt{1 - v^2/c^2}$. Therefore, in the frame of the bee, the distance travelled in going from one end of the stick to the other is only $L = vt_{bee}$, i.e.,

$$L = L_0/\gamma. \tag{27}$$

Of course, for real bees, the factor γ is so close to 1 that t_{bee} and t_0 are for all practical purposes indistinguishable. However, subatomic particles are often accelerated in high energy physics laboratories to speeds very close to c, so that $\gamma \gg 1$. For example at Stanford University, electrons are accelerated in a 3-km (approximately 2-mile)-long straight machine called a linac. By the time an electron gets to the end of the linac, its speed is so close to c that the Lorentz factor $\gamma \approx 10^5$. In the rest frame of an electron with this speed, the apparent length of the linac is only about three centimeters (little more than an inch!).

B. Velocity Combination Formula

Consider two inertial frames such that the velocity of one frame (the primed frame) with respect to the other frame (the unprimed frame) is v along the x-axis. We now suppose there is an object having velocity V', also along the x-axis, as observed in the primed frame. What is the velocity V of the object as observed in the unprimed frame? Commonsense experience would lead us to expect that the velocity in the unprimed frame is just $V = v + V'$. As shown earlier, this is what the Galilean transformation gives. However, relativity tells us that the correct transformation between frames is the Lorentz transformation, which gives

$$\begin{aligned} dx &= \gamma(dx' + v\,dt') \\ dt &= \gamma\left(dt' + \frac{v}{c^2}\,dx'\right). \end{aligned} \tag{28}$$

Dividing the first equation by the second, and using the fact that $V' = dx'/dt'$ and $V = dx/dt$, we obtain

$$V = \frac{v + V'}{1 + vV'/c^2}. \tag{29}$$

The closer the velocities are to the speed of light, the more this expression disagrees with simple addition of velocities. Note that if either of the velocities v or V' is equal to c, then V is equal to c. So this is consistent with the postulate that the speed of light *in vacuum* looks the same in all reference frames. We also see that so long as both v and V' are less than c, the magnitude V of the combined velocity will also be less than c. More general velocity combination formulas, where the velocities are not all along the same axis, may similarly be derived from the Lorentz transformation, and these conclusions still hold.

C. Relativistic Doppler Effect

It is familiar from our everyday experience that the pitch of a siren is higher when it approaches and lower when it recedes. This is an example of the Doppler effect, which is a change in the observed frequency of a periodic disturbance, arising from the motion of the source and/or the observer. In the case of sound waves, the observed speed of the waves, as well as the observed frequency, depends on the motion of the observer with respect to the transmitting medium. In other words, sound waves have a "preferred" frame, namely, the frame in which the transmitting medium, typically air, is at rest. Light and other electromagnetic waves, however, do not have a preferred frame—we have seen that attempts to find an "ether" failed.

For comparison, we will first derive the nonrelativistic Doppler formula for sound waves. Then we will derive the relativistic Doppler effect for light, i.e., for electromagnetic waves. Throughout the discussion we use the general relationship

$$V = \lambda f \tag{30}$$

between the propagation speed V, wavelength λ, and frequency f of a wave or other periodic disturbance. Also, the frequency of the wave is just the reciprocal of its period T:

$$f = 1/T. \tag{31}$$

1. Nonrelativistic Doppler Effect for Sound Waves

The speed of sound is much less than c, so if we assume the speeds of the source and observer are also much less than c, then Newtonian mechanics is valid to an excellent approximation. Suppose that a source of sound waves S and a receiver R are moving along the same line which we take to be the x-axis, and suppose that their velocities with respect to the air are v_S and v_R, respectively (we speak of velocities v_R and v_S here rather than just speeds, since these quantities do have a sign according to whether they are in the direction of the positive or negative x-axis). Let

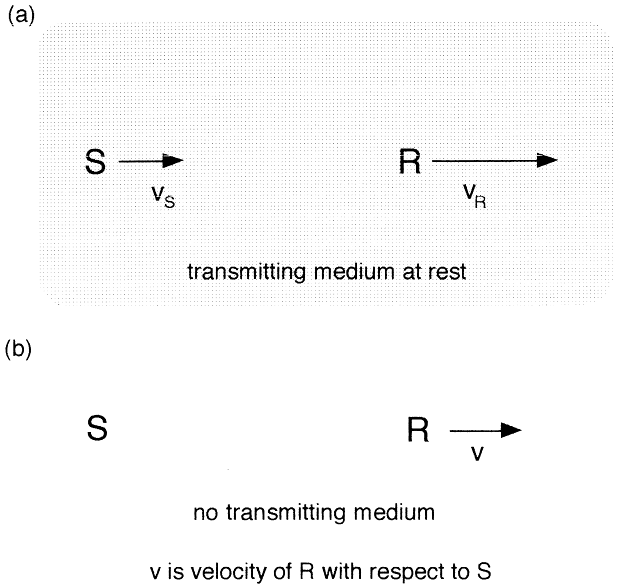

(a)
S
v_S
R
v_R
transmitting medium at rest
(b)
S
R
v
no transmitting medium
v is velocity of R with respect to S

FIGURE 6 (a) Diagram for analyzing nonrelativistic Doppler effect for sound waves. The effect depends on the velocities of the source S and the receiver R with respect to the transmitting medium (typically air). (b) Diagram for analyzing relativistic Doppler effect for electromagnetic waves. The effect depends only on the relative velocity of the source S (whose rest frame is the unprimed frame) and receiver R (whose rest frame is the primed frame).

R be to the right of S, as shown in Fig. 6(a). Let S be emitting a sound wave of frequency f (and thus period $T = 1/f$), and suppose that the speed of sound in the air is w. The time between emission of successive crests of the wave is $T = 1/f$. The distance between crests of the wave would be $\lambda = w/f$ if the source were at rest. However, the velocity of these crests relative to the source is $w - v_S$. so for nonzero v_S the distance between crests is modified to

$$\lambda_{mod} = (w - v_S)/f. \qquad (32)$$

If R were at rest, the time between the arrivals of successive crests would be λ_{mod}/w. But for nonzero v_R, the velocity of the wave with respect to R is $w - v_R$, so that the time between arrivals of successive crests is $T_{mod} = \lambda_{mod}/(w - v_R)$. The received frequency f_{mod} is just $1/T_{mod}$, so that we end up with

$$f_{mod} = f\frac{1 - v_R/w}{1 - v_S/w}. \qquad (33)$$

This formula relates the frequency f emitted by S and the frequency f_{mod} received by R. The velocities v_R and v_S are positive toward the right and negative toward the left. Note that the effect depends on both v_R and v_S and cannot be reduced to an expression involving only their relative velocity v_R–v_S.

2. Relativistic Doppler Effect for Electromagnetic Waves

We now analyze the relativistic Doppler effect for electromagnetic waves. We again assume that the velocities of the source and observer are along the x-axis, and we choose the unprimed frame to be the rest frame of the source and the primed frame to be the rest frame of the receiver [see Fig. 6(b)].Then the velocity v of the primed frame with respect to the unprimed frame is just the relative velocity of the source and receiver. In the frame of the source, let the period of the wave be T, i.e., T is the time between emission of wave crests as seen in the source frame. The wavelength in the source frame is then $\lambda = cT$.

Now let us consider two events and how they look in the two frames. Let Event 1 be the arrival of the crest of the nth wave at the receiver, and let Event 2 be the arrival of the crest of the $n + 1$st wave at the receiver. From the point of view of the source frame, the velocity of the wave train with respect to the receiver is $c - v$. Thus, in the source frame, the time separation between Event 1 and Event 2 is

$$\Delta t \equiv t_2 - t_1 = \frac{\lambda}{c - v} = \frac{cT}{c - v}. \qquad (34)$$

Since the receiver is traveling at velocity v in the source frame, the distance between the two events in the source frame is

$$\Delta x \equiv x_2 - x_1 = v\Delta t = \frac{vcT}{c - v}. \qquad (35)$$

Now we need only Lorentz transform from the source frame (the unprimed frame) to the receiver's frame (the primed frame) to find the time separation between the two events *as observed in frame of the receiver*. This is just the period of the waves as seen by the receiver:

$$T_{mod} = \Delta t' = \gamma\left(\Delta t - \frac{v}{c^2}\Delta x\right). \qquad (36)$$

Substituting for Δt and Δx from Eqs. (34) and (35), we find

$$T_{mod} = \gamma\left(\frac{cT}{c - v} - \frac{v}{c^2}\frac{vcT}{c - v}\right) = \frac{\gamma cT}{c - v}\left(1 - \frac{v^2}{c^2}\right). \quad (37)$$

Using the usual definitions $\beta = v/c$, $\gamma = 1/\sqrt{1 - \beta^2}$, and $f = 1/T$ as well as $f_{mod} = 1/T_{mod}$ we may simplify this to

$$f_{mod} = \left(\frac{1 - \beta}{1 + \beta}\right)^{1/2} f. \qquad (38)$$

Note that v (and thus β) is positive if the source and receiver are moving away from each other, and is negative if the source and receiver are moving toward each other. Since there is no "preferred frame" as there was in case of sound waves, the result depends only on the relative velocity of the source and receiver.

D. Relativity of Simultaneity: Einstein's Train Paradox

It is necessary to be extremely careful about using the phrase "at the same time" when dealing with relativity. The Newtonian and Galilean concept of an absolute time, flowing uniformly and at the same rate for all observers, is not strictly true. The notion of an absolute time is an excellent approximation in our everyday experience. However, it breaks down when considering situations involving relative motion at very high speed.

To see this, we examine a situation sometimes referred to as "Einstein's train paradox." Consider a (very high-speed!) train moving along a straight track. Let a woman be at sitting on the train exactly equidistant from its two ends, and let a man be standing on the ground right next to the tracks. Suppose that the man sees the flashes from two bolts of lightning striking the ends of the train at the exact same moment when the woman is passing him on the train. At this moment the man knows he is also equidistant from both ends of the train (from symmetry, assuming he knows the woman is sitting equidistant from both ends). So, knowing that the speed of light is c, he concludes that the light flashes from each end will take the same amount of time to reach him and therefore the strikes at each end of the train occurred simultaneously. He also concludes that the flash from the front of the train will arrive at the woman before the flash from the back arrives, because she is moving toward the front flash and away from the back flash.

Now, what does the woman say about these events? The mere fact of changing reference frames, e.g., from the man's frame to the woman's frame, *cannot change the time ordering of events which occur at the same location in some frame* (in this case, at the location of the woman in her frame). To see this, suppose for example that the woman is carrying a device which does nothing if it receives the front flash first, and explodes if it receives the back flash first. Obviously, the functioning of the device must be independent of whose frame the situation is analyzed from—the explosion either occurs or does not occur in both frames. So the device must receive the flashes in the same order in both frames.

At first glance there is nothing surprising about the fact that the woman sees the front flash before the back flash. Based on our ordinary experience with velocities we would be tempted to say that in the woman's reference frame, the front flash travels faster than c and the back flash travels slower than c, if c is the speed of light in the man's frame. However, one of the postulates of relativity is that the speed of light is observed to be the same in *any* frame, regardless of the state of motion of the observer. So the front and back flashes also both travel with speed c in the woman's frame. A consequence of this is that the two observers do not agree about whether the two flashes are simultaneous.

This may seem paradoxical at first, because it is not something we are familiar with from everyday experience. We see that accepting Postulate 2 (invariance of the speed of light) forces us to revise the concept of simultaneity. *Relativity of simultaneity* applies to any two events which are separated along the line of relative motion of two different inertial frames: If the two events are simultaneous in one of the frames, they will not be simultaneous in the other—this follows directly from the Lorentz transformation. The train "paradox" is just one illustration of this general statement.

Most "paradoxes" in relativity can in the end be reduced to some confusion arising from the fact that whether or not two events are simultaneous depends on the observer's frame of reference. With this in mind, it is instructive to look at two more famous relativistic paradoxes and how they are resolved.

E. The Twin Paradox

According to the principle of relativity, all processes in a given frame, including the biological workings of a human body, must undergo the same amount of time dilation as all the other clocks and physical processes in that frame. Any inertial frame is supposed to follow the same physical laws as any other inertial frame, and this would be violated if a person saw a clock in his own frame run slower with respect to his own heartbeat when he changed his speed (of course, if this happens to a race car driver, it is a psychological and not a relativistic effect!)

A very famous paradox involving relativistic time dilation is the twin paradox. Suppose there are twins, Astro and Homer, and that Astro makes a round trip journey to a star 20 light-years away while Homer remains on the Earth. A light-year is the distance light travels in 1 year. [Here we use the convenient device of measuring time and distance in the same unit, the year.] To make the analysis simple, we assume that Earth and the destination star are both at rest in the same inertial frame, and that Astro travels at constant speed (say 80% of the speed of light) straight to the star, instantaneously reverses direction upon getting there, and returns at the same speed. Since the one-way distance L is 20 light-years in Homer's rest frame and $\beta = v/c = 0.8$, Homer calculates that the total time T for the trip will be $2L/v = 50$ years.

However, from the Lorentz time dilation, Homer knows that Astro's clock will be running slow by the factor

$\gamma = [1-(0.8)^2]^{-1/2} = 5/3 \approx 1.67$ compared to his own, on both the outbound and inbound parts of the trip. So he would expect Astro to have experienced an elapsed time of $T' = T/\gamma = 50/(5/3) = 30$ years, and indeed this is what would happen. Another way to see this is to note that from Astro's point of view, the one-way distance is Lorentz contracted to $L' = L/\gamma = 20/(5/3) = 12$ light-years. Astro sees Earth and the star speeding past himself at speed $v = 0.8c$, so according to him the journey takes $T' = 2L'/v = 30$ years. The net effect is that when Astro returns he is 20 years younger than Homer.

The "paradox" consists of the fact that in accord with Lorentz time dilation, Astro should also expect Homer's clock to be running slow relative to his own, by the same factor γ, since Homer is in motion with respect to Astro at the same speed v. So, exactly how does the difference in their ages come about?

We already know the answer if we recall our discussion of proper time. We saw that the time between two events is maximized for a clock which goes from one event to another by staying in the same inertial frame all the way. It is true that while each twin is in an inertial frame, and thus moving at constant velocity with respect to other twin, each will observe the other's clock to be ticking more slowly than his own. The difference is that Homer remains in the same inertial frame throughout, while Astro changes inertial frames when he reverses direction. There are real physical effects felt by Astro when this acceleration occurs—he is dragged toward one end of the ship as it turns around (in fact the extreme acceleration assumed in this example would kill a real human!).

The change of inertial frames is associated with a shift in Astro's definition of simultaneity at locations away from him along his line of motion. In particular, his definition of which events on earth are simultaneous with events ocurring at his location will abruptly shift when he changes reference frames. To see this, let us denote the rest frame of Homer by S, the rest frame of Astro on the outbound leg by S', and the rest frame of Astro on the inbound leg by S''. Take frame S' to be moving with speed v in the positive x direction with respect to S, and take frame S'' to be moving with speed v in the negative x direction with respect to S. A space–time diagram of the trip as viewed in the S frame is shown in Fig. 7. The solid line is Astro's world line—he travels 20 light-years in an elapsed time of 25 years, then reverses direction and returns to his starting point in another 25 years of elapsed time. In Fig. 7, all events lying on any given horizontal line are simultaneous in the S frame since they have the same t coordinate. But observers in frames moving with respect to the S frame have different definitions of simultaneity. If an observer is moving along the x direction of frame S, his lines of simultaneity will not be horizontal in Fig. 7.

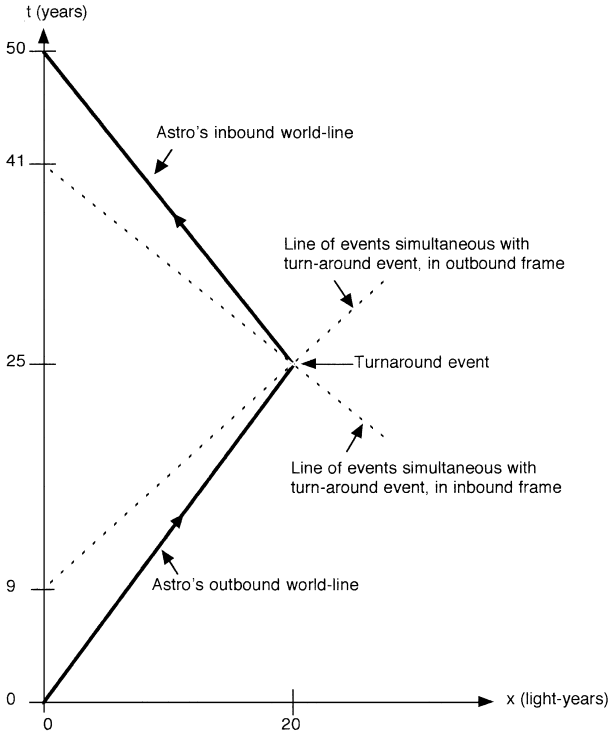

t (years)
50
41
25
9
0
0
20
x (light-years)
Astro's inbound world-line
Line of events simultaneous with turn-around event, in outbound frame
Turnaround event
Line of events simultaneous with turn-around event, in inbound frame
Astro's outbound world-line

FIGURE 7 Spacetime diagram of Astro's journey, as seen in the rest frame of Homer. The heavy line is Astro's world-line. The dotted lines show the lines of simultaneity with the turnaround point, in Astro's outgoing and ingoing frames.

The upward-sloping dotted line in Fig. 7 is the line containing events simultaneous with the turnaround event, *according to observers in frame S'*. We may easily calculate the slope of this line by applying the Lorentz transformation. Let t_1, x_1 and t_2, x_2 be the coordinates of any two events in the S frame. Then the time coordinates in the S' frame are given by the Lorentz transformation, which, with time and distance measured in the same units, becomes

$$t'_1 = \gamma(t_1 - \beta x_1) \tag{39}$$

$$t'_2 = \gamma(t_2 - \beta x_2). \tag{40}$$

so that

$$\Delta t' = \gamma\left(\Delta t - \frac{v}{c}\Delta x\right), \tag{41}$$

where $\Delta t \equiv t_1 - t_2$, $\Delta x \equiv x_1 - x_2$, and $\Delta t' \equiv t'_1 - t'_2$. If the two events are simultaneous in S' (i.e., $\Delta t' = 0$), then $\Delta t/\Delta x = \beta = v/c$. Thus all lines of simultaneity in frame S', the frame of Astro during his outbound journey, have slope v/c in Fig. 7. We show one such line, in particular, the line of all events simultaneous with the turnaround event, according to Astro while he is in the outbound frame.

This line of simultaneity for the outbound frame S' intersects the t-axis, where $x = 0$ (i.e., at the location of Earth), at $t = 9$ years. This is in agreement with the assertion that Astro sees Homer's clock run slow by the factor γ—the outbound trip takes 15 years according to Astro, thus $15/\gamma = 9$ years should elapse on Homer's clock, according to Astro's definition of simultaneity.

By similar reasoning, the lines of simultaneity for frame S'', the frame of Astro during his inbound journey, have slope $-v/c$. The downward-sloping dotted line is the line of all events simultaneous with the turnaround event, according to Astro when he is in the inbound frame. The intercept on the t-axis is at $t = 41 = 50 - 9$ years, again consistent with Astro's expectation that 9 years should elapse on Homer's clock during Astro's return journey. However, Astro's definition of simultaneity with events on Earth jumps ahead by $41 - 9 = 32$ years when he changes frames! Of course, no sudden jump occurs on the clocks in any frame! It is just that the definition of simultaneity of events is relative—it depends on the observer's frame of reference. Clocks at a given location (say at the earth) which have been synchronized in different frames will in general have different readings. Suppose that the twins were exactly 30 years old when Astro departed on his journey. If Astro has just reached the turnaround point, he will say that it is his own 45th birthday and if he is still in the outbound frame, he will say that back on Earth it is "now" Homer's 39th birthday. However, if an instant later he has made the transition to the inbound frame, he will say that Homer is "now" celebrating his 71st birthday. According to Astro's clocks it takes 15 years to return, making him 60 years old at the twins reunion. As noted before, according to Astro, $15/\gamma = 9$ years will elapse for Homer during the return journey, making Homer 80 years old at their reunion.

We can gain further insight by looking at how the situation evolves if each twin is periodically sending out a light flash at some frequency f as measured by his own clock, say one flash at each birthday. Then each twin can count the flashes received from the other, and by the time the twins come back together again, each twin must have received all the signals sent out by the other during the trip. To be consistent with Homer's faster aging, Astro should receive a total of 20 more flashes than Homer does over the course of the trip.

First consider what Homer sees. He says each leg of the trip takes 25 years. Thus the flash from the turnaround event originates 25 years after Astro's departure from Earth, according to Homer. However, he doesn't see it until 45 years after Astro's departure from Earth because the turnaround point is 20 light-years away and so this flash takes another 20 years to get back to Earth. According to the Doppler formula, Homer receives pulses originating during the outbound leg at the rate

$$\left(\frac{1-\beta}{1+\beta}\right)^{1/2} \cdot (1 \text{ flash per year}) = 1/3 \text{ flash per year}. \tag{42}$$

Therefore he receives a total of

$$(45 \text{ years})(1/3 \text{ flash per year}) = 15 \text{ flashes}. \tag{43}$$

During the inbound leg he receives pulses at the rate

$$\left(\frac{1+\beta}{1-\beta}\right)^{1/2} \cdot (1 \text{ flash per year}) = 3 \text{ flashes per year}. \tag{44}$$

He receives these flashes over the course of the remaining 5 years, for a total of $5 \cdot 3 = 15$ more flashes. Thus he has received a total of $15 + 15 = 30$ flashes, as he should since Astro ages 30 years and therefore sent Homer 30 flashes during the trip.

Now let us calculate how many flashes Astro sees. According to him the outbound leg of the trip takes 15 years, during which time he receives flashes at the rate of $1/3$ per year, for a total of 5 flashes. The inbound trip also takes 15 years, during which time he receives flashes at the rate of 3 per year, for a total of 45 flashes. Thus over the course of the entire trip he receives $45 + 5 = 50$ flashes, in agreement with the fact that Homer has aged 50 years when they are reunited.

If a real human were going on such a journey, he or she would need to be accelerated gradually rather than instantaneously. The calculation of the age difference of the twins, although more complicated, could still be done and would still show Astro to end up younger than Homer. Our civilization does not yet have the technology and resources to send living beings on such high-speed journeys, so this particular experimental test has not been done! However, time-dilation effects have been experimentally verified by observing subatomic particles. For example, there is a type of particle called a muon, which has only a short lifetime before it disappears, producing other particles. (The produced particles are the familiar electron and particles called neutrinos.)

It cannot be predicted exactly when this "decay" of a given muon will occur, but the *average* lifetime before decay is about 2×10^{-6} sec. This is the lifetime of a muon as observed in its rest frame. However, muons often travel at speeds very close to the speed of light relative to observers on the Earth. When a large number of muons having a given speed are observed, it is found that the average of their lifetimes does indeed increase by the factor γ.

Time dilation is the reason that significant numbers of muons from cosmic rays are observed at the Earth's surface. Muons are produced high in the atmosphere of the

Earth when cosmic rays enter from outer space.If their lifetimes were not increased due to travelling near the speed of light, very few of them would reach the surface of the Earth before decaying.

F. The Pole and Barn Paradox

Another famous paradox has to do with Lorentz contraction of moving objects. Suppose that there are a long pole and a barn which both have proper length L. The barn has a door at each end that may be quickly opened and closed. An unbelievably fast runner carries the pole and runs through the barn with it, with the pole horizontal and pointed in the runner's direction of motion. In the rest frame of the barn, the pole is moving and so it would appear shorter than the barn. Therefore the pole should easily fit into the barn. It should be possible to close both doors for a short time, with the pole entirely inside the barn. But in the rest frame of the pole, it is the barn that is moving, and so it would appear shorter than the pole. Therefore the pole should be too long to fit completely inside the barn at any time. At first glance, it seems that we have a contradiction.

The resolution of course has to do with the differing definitions of simultaneity in the two frames. To be definite, suppose that the pole is being carried so fast that the Lorentz factor $\gamma = 2$. Then the moving pole, as seen in the rest frame of the barn, is contracted to half its proper length. [This would mean that in the barn frame $\beta \equiv v/c = \sqrt{3}/2$, i.e., the pole is moving at about 87% of the speed of light.]

A spacetime diagram plotted in the frame of the barn (which we take to be the unprimed frame) is shown in Fig. 8(a). The world lines of the front door of the barn (labeled F), the back door of the barn (labeled B), the head of the pole (labeled H), and the tail of the pole (labeled T) are straight lines as shown. We assume that the front door is closed immediately after the tail of the pole has passed through it. The world line of the front door is cross-hatched while the front door is closed. We also assume that the back door is opened immediately before the head of the pole passes through it. The world line of the back door is cross-hatched while the back door is closed.

We label four events as follows:

1. Event HF. Head of pole passes through front door of barn
2. Event TF. Tail of pole passes through front door of barn
3. Event HB. Head of pole passes through back door of barn
4. Event TB. Tail of pole passes through back door of barn

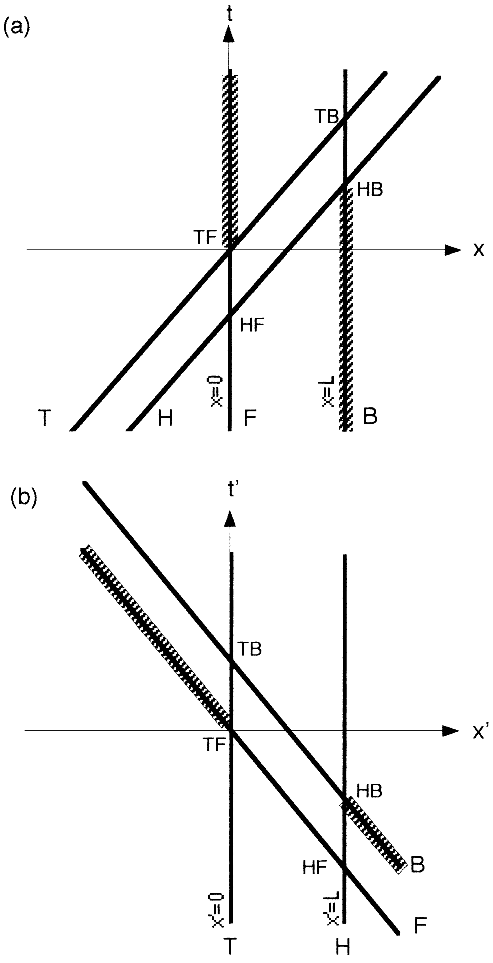


FIGURE 8 (a) Spacetime diagram in rest frame of barn. (b) Spacetime diagram in rest frame of pole. H labels world-line of head of pole, T labels world-line of tail of pole, F labels world-line of front of barn, and B labels world-line of back of barn. Event HF = "head of pole passes through front door of barn," event TF = "tail of pole passes through front door of barn," event HB = "head of pole passes through back door of barn," event TB = "tail of pole passes through back door of barn." Cross-hatching on the world line of a door means the door is closed.

We have chosen the origin of coordinates to be the at event TF. Figure 8(a) shows that in the frame of the barn, the four events take place in the order in which we have just listed them. In this frame both doors are closed during the time interval between TF and HB. This is fine, since in this frame the pole is only half as long as the barn, so it easily fits inside for the period of time between events TF and HB.

Figure 8(b) depicts the same situation plotted in the rest frame of the pole (the primed frame, moving at velocity $v = \sqrt{3}c/2$ with respect to the unprimed frame. This figure shows that the time order of the events TF and HB is different in the two frames! In the pole's rest frame, the back door is opened, allowing the head of the pole to pass through *before* the tail of the pole passes through the front door. In this frame, there is never a time when both doors are simultaneously closed. There had better not be, because in this frame the pole is twice as long as the barn, so it cannot fit completely inside at any time. In neither frame does the pole touch either door. Something would be wrong if our reasoning told us the pole could dent a door in one frame but not in the other. We could look at the doors after the fact, and whether or not there is a dent cannot depend on the observer's frame of reference. However, the answer to the question "Was the pole ever entirely inside the barn?" *does* depend on the observer's frame of reference.

Suppose that an observer in the barn rest frame decides she will "prove" that the pole is "really" shorter than the barn by suddenly bringing it to rest while (according to her) it is totally inside the barn. However, "suddenly bringing the pole to rest" means that all pieces of it are brought to rest simultaneously. As we already know, events along the direction of motion—in this case along the length of the pole—that are simultaneous in the barn frame are not simultaneous in the pole frame. Furthermore, no causal influence can go faster than c. It would not be sufficient for the barn-frame observer to just grab the pole at its center with a single short clamp. There is no immediate effect on the rest of the pole—it takes a nonzero amount of time for the ends of the pole to even "know" that the clamp has taken hold of the middle, since no causal influence can go faster than c. Thus we need many clamps stationed along the path of the pole if we want to stop all parts of the pole simultaneously.

So, let us assume that the observer in the barn rest frame times her clamps so that all parts of the pole stop at the same time in her frame. When the pole has been brought to rest in the barn, it has been crushed to a length of $L/2$. This is now the proper length of the pole since the pole is now at rest in the barn frame. The structure of the pole really has been changed—the barn observer has crushed the pole to half its original proper length using her stopping method.

How does this look to an observer in the original rest frame of the pole? His frame continues to move at velocity $v = \sqrt{3}c/2$ with respect to the barn. According to him, clamps near the head of the pole take hold before clamps near the tail of the pole. So he agrees that the pole will get crushed to a shorter length. But the final length as observed in his frame is not its proper length, since the pole does not end up at rest with respect to his frame. He continues to move at speed v with respect to the barn frame, which is the new rest frame of the pole. Therefore the pole is Lorentz contracted from its new proper length of $L/2$ by another factor of $\gamma = 2$, to an apparent length of $L/4$ in his frame.

Other stopping methods are of course possible. All segments of the pole could be brought to rest with respect to the barn by stopping each piece simultaneously *as measured in the initial rest frame of the pole*. This means that each segment of the pole is given a backwards velocity of $-v$ as observed in the initial rest frame of the pole. The length of the pole in the initial pole rest frame would still be L but this would no longer be its proper length since it ends up moving at $-v$ with respect to that frame, and hence is Lorentz-contracted by the factor $\gamma = 2$. Once again, the pole has been deformed to a new proper length—this time it is stretched or pulled apart to a new proper length of $\gamma L = 2L$.

Stopping the pole while maintaining its original proper length throughout the transition from the original pole rest frame to the barn rest frame is more complicated to analyze. In this case, the pole moves smoothly through an infinite sequence of different Lorentz frames. Einstein's general theory of relativity is better suited to handling such situations involving continuous acceleration than is special relativity.

Obviously, runners carrying poles do not really run fast enough for relativistic effects to be significant! But the electrons in the Stanford linac typically travel down the linac in groups ("bunches") that are about a centimeter long as observed in the rest frame of the linac. For a Lorentz factor $\gamma = 10^5$, the proper length of a bunch (that is, the length as observed in the rest frame of the bunch) is therefore about a kilometer. As noted earlier, in the frame of an electron with this value of γ the linac is only about three centimeters long, obviously much shorter than the proper length of the bunch. Whether or not the bunch "fits" inside the linac depends on your reference frame!

V. RELATIVISTIC TREATMENT OF ENERGY AND MOMENTUM

Two very fundamental laws of mechanics are conservation of energy and momentum. These laws are extremely useful in analyzing physical situations. They say that the total amount of energy and the total amount of momentum in an isolated physical system do not change with time. Since they are so useful, we would hope to have similar conservation laws in special relativity. We will also want the new relativistic definitions of mass, energy, and momentum to reduce to the old Newtonian ones in situations where the velocities involved are much less than the speed of light. All this can be done, but some modifications to our notions of mass, energy, and momentum are required. To motivate the new definitions, we will look at situations involving collisions between two objects. Even in these simple situations the arguments are a little more complicated than in most of this article. However, following through the reasoning will yield some further insight into the analysis of relativistic problems. The main results are summarized at the end of this section. The usefulness and validity of the new definitions have been borne out by all known relevant experimental evidence.

A. Relativistic Momentum

Momentum conservation would *not* be preserved under Lorentz transformations, if we used the Newtonian definition of momentum $\mathbf{p} = m\mathbf{v}$, where m is the mass. Let us try introducing a multiplicative function $f(v)$ that is a function of the speed v of the object (remembering that speed v is just the magnitude of the velocity $\mathbf{v}$, i.e., $v = |\mathbf{v}|$). We start by guessing that the momentum in a frame where the object moves with velocity $\mathbf{v}$ is given by

$$m f(v)\mathbf{v}. \tag{45}$$

Our first goal is to see if we can find a functional form for the dependence of $f(v)$ that satisfies both momentum conservation and the principle of relativity (i.e., momentum conservation holds in all inertial frames). For $v \ll c$ we want the momentum to agree with the Newtonian definition, thus we require $f(v) \to 1$ as $v \to 0$.

To find out what $f(v)$ must be, we consider an elastic collision between two objects of equal mass m. "Elastic" means that the two objects bounce apart without dissipating any of their incoming kinetic energy as other forms of energy. We will look at this collision from two different inertial frames. Choose one frame to be the center of mass (CM) frame, which is the frame in which the total momentum is zero. In this frame the momenta of A and B are equal in magnitude and opposite in direction. We can choose the spatial coordinate axes in this frame, which we will label the unprimed frame, so that the collision looks as shown in Fig. 9(a). [Note this is not a diagram in space–time; it is simply a diagram of the paths of the objects in two space dimensions.] Since the masses are equal and the collision is symmetric in this frame, A has an equal and opposite velocity to B, both before and after the collision. Since the collision is elastic, the speed of each object before the collision is the same as its speed afterwards.

In this frame, we assume that A departs from the line $y = -D/2$ at the same time B departs from the line $y = +D/2$, and they have equal speeds v. At a time $T_{CM}/2$ later, they collide at the origin and each reverses its component of velocity along the y-axis. After an additional time $T_{CM}/2$, A returns to the line $y = -D/2$ and B returns to the line $y = +D/2$. Since the total distance travelled by each object along the y direction during the time T_{CM} is D, the magnitude of the y velocity component is D/T_{CM}. Since each object reverses the direction of its y velocity component during the collision, the y component of momentum change in the collision is $-2mf(v)D/T_{CM}$ for A and $+2mf(v)D/T_{CM}$ for B. From the symmetry of the collision as viewed in this frame, it is obvious that the total momentum is conserved—it is zero both before and after the collision.

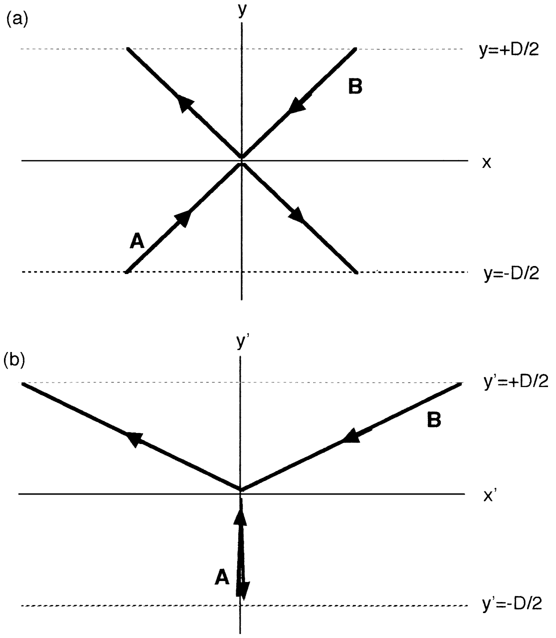


FIGURE 9 Elastic collision of particles A and B. (a) As viewed in center of mass frame (chosen to be unprimed frame). (b) As viewed in frame moving with longitudinal velocity of particle A (primed frame).

Now we impose the requirement that momentum conservation should hold for any other inertial frame as well. In particular, we consider a (primed) frame that is moving to the right at speed v_x with respect to the unprimed frame. We choose v_x to be the velocity component of A along the x-axis, as observed in the unprimed frame. Thus in the primed frame A just moves straight up the y axis and back down, while B follows a longer path, as shown in Fig. 9(b), with velocity component $v'_{x,B}$ along the x-axis. [Note that $v'_{x,B}$ is NOT the value $-2v_x$ that we would get if the Galilean transformation, and hence straightforward addition of velocities, were valid.]

We know that Lorentz transformations along a given direction do not change the observed distances transverse to that direction, so the total distance in the y direction travelled by each object is also D in the primed frame. We also know that events which are simultaneous in one frame are not simultaneous in another, when the events are separated along the direction of relative motion of the two frames. By symmetry, the departure events were simultaneous in the unprimed frame and so were the return events; we implicitly assumed this when we said that the total time of the trip for each object was the same value T_{CM}. But this is not true in the primed frame, contrary to what nonrelativistic intuition would lead us to expect. The departure and return events for A occur at the same place in the primed frame. So, the time between these events as measured in the primed frame is the proper time between the events, which we will call T_0. Therefore, as observed in the primed frame, the momentum change in the collision for A is

$$\Delta p'_A = -2mf(v'_A)D/T_0, \tag{46}$$

where $v'_A = D/T_0$ is the total speed of A in the primed frame.

Now let us calculate the momentum change of B during the collision, as observed in the primed frame. If we had been in a frame in which B moved straight down and then back up the y-axis, then by symmetry B's round trip time would have been T_0 in that frame. But the primed frame moves at velocity $v'_{x,B}$ with respect to such a frame. Therefore, from Lorentz time dilation, the time between the departure and return of B in the primed frame is

$$T = \frac{T_0}{\sqrt{1-(v'_{x,B}/c)^2}}. \tag{47}$$

So, as observed in the primed frame, the momentum change in the collision for B is

$$\Delta p'_B = 2mf(v'_B)D/T. \tag{48}$$

where $v'_B = \sqrt{(v'_{x,B})^2 + (D/T)^2}$ is the total speed of B in the primed frame.

The momentum changes in Eqs. (46) and (48) must sum to zero if conservation of momentum is to hold in the primed frame. Using Eq.(47) to eliminate T/T_0, this requirement gives

$$mf(v'_A) = mf(v'_B)\sqrt{1-(v'_{x,B}/c)^2}. \tag{49}$$

The final step in this argument is to take the limit of this expression as D approaches zero, so that the y-components of the velocities of both objects approach zero. In other words, we consider the extreme case where A and B just graze each other. In this limit, A is at rest in the primed frame so that $f(v'_A) \rightarrow f(0) = 1$. Also, in this limit B's velocity is entirely along the x-axis, so that $v'_{x,B} = v'_B$ is its total speed. Thus we obtain

$$mf(v'_B) = \frac{m}{\sqrt{1-(v'_B/c)^2}}. \tag{50}$$

This example shows that our initial assumption [that momentum is given by $mf(v)\mathbf{v}$] can only hold if the functional form for $f(v)$ is

$$f(v) = \frac{1}{\sqrt{1-v^2/c^2}}, \tag{51}$$

where v is the speed. Thus $f(v)$ is just the Lorentz factor γ introduced previously. The resulting definition of momentum is

$$\mathbf{p} = \gamma m\mathbf{v} = \gamma m\, d\mathbf{x}/dt. \tag{52}$$

This prescription gives consistent results in agreement with experiment and is what is adopted in special relativity.

B. Relativistic Energy and the Momentum-Energy Four-Vector

Note that the relativistic definition of momentum, Eq. (52), can also be expressed as

$$\mathbf{p} = m\, d\mathbf{x}/d\tau. \tag{53}$$

Here we simply differentiate with respect to proper time along the world line of the object.

Given this way of writing **p**, it is natural to try adding a fourth component

$$p_t = md(ct)/d\tau = mc\gamma \tag{54}$$

and see whether the resulting four-vector is a useful concept. Consider the quantity analogous to the space–time interval, but replacing t, x, y, z by p_t, p_x, p_y, p_z:

$$\begin{aligned} p_t^2 - p_x^2 - p_y^2 - p_z^2 &= m^2c^2\left(\frac{dt}{d\tau}\right)^2 - m^2\left(\frac{dx}{d\tau}\right)^2 \\ &\quad - m^2\left(\frac{dy}{d\tau}\right)^2 - m^2\left(\frac{dz}{d\tau}\right)^2 \\ &= m^2\frac{[c^2\,dt^2 - dx^2 - dy^2 - dz^2]}{d\tau^2}. \end{aligned} \tag{55}$$

The expression in brackets in the last line is just the space–time interval ds^2. Assuming that this interval is timelike (as it must be if the object is traveling at less than the speed of light), we have $ds^2 = c^2\,d\tau^2$ so we are left with

$$p_t^2 - p_x^2 - p_y^2 - p_z^2 = m^2c^2. \tag{56}$$

Like the space–time interval, this quantity is an invariant since it depends only on the mass m and the speed of light c, both of which are the same in all frames. But what is the real significance of p_t? We have already concluded that the momentum vector (p_x, p_y, p_z) is conserved. It therefore seems natural to guess that p_t might also be conserved and related to the energy. Multiplying p_t by another factor of c to get energy in the usual units, we make the guess that γmc^2 is the energy of an object with mass m and speed v.

If we define the total energy in this way, we see that there is a new contribution to the energy. Even when the speed of the object is zero, it still has an nonzero energy, the *rest energy* given by

$$E_{rest} = mc^2, \tag{57}$$

where m is the object's mass. As stated earlier, we define the energy of a moving object to be

$$E_{tot} = \gamma mc^2, \tag{58}$$

where $\gamma = 1/\sqrt{1 - v^2/c^2}$ and v is the object's speed.

To see that this definition of the energy reduces to something reasonable in the nonrelativistic limit $v \ll c$, we expand it in a power series in v/c:

$$E = \gamma mc^2 = mc^2\left[1 - \frac{v^2}{c^2}\right]^{-1/2} = mc^2\left[1 + \frac{1}{2}\frac{v^2}{c^2} + \cdots\right]$$
$$\approx mc^2 + \frac{1}{2}mv^2 + \cdots \tag{59}$$

Thus in this limit the energy is the sum of the rest energy, the nonrelativistic kinetic energy $\frac{1}{2}mv^2$, and higher order terms in v^2/c^2 that are much smaller than the first two terms when $v \ll c$.

Let us verify that energy defined in this way is conserved in an inelastic collision, namely one in which two objects collide head-on and stick together. Suppose the two objects have equal masses m and we view the collision in a frame (call it the primed frame) where the objects approach each other with equal speeds v along opposing directions, say parallel to the x-axis [see Fig. 10(a)]. Since the masses of the two incoming objects are equal, this is the center-of-mass (CM) frame, i.e., the frame in which the total momentum of the system is zero. From momentum conservation, the objects will be at rest in this system after they collide and stick together.

We have chosen A to be the object moving to the right and B the one moving to the left. Now let us Lorentz transform to the rest frame of B; call this the unprimed frame. The primed frame moves with speed v to the right with respect to the unprimed frame. The collision as observed in the unprimed frame is shown in Fig. 10(b). From the velocity combination formula, the velocity of A in the unprimed frame is

$$V = \frac{2v}{1 + v^2/c^2}. \tag{60}$$

A little algebra then gives

$$1 - V^2/c^2 = \frac{(1 - v^2/c^2)^2}{(1 + v^2/c^2)^2}. \tag{61}$$

Therefore the energy of A in the unprimed frame is

$$E_A = \frac{mc^2}{\sqrt{1 - V^2/c^2}} = mc^2\frac{(1 + v^2/c^2)}{(1 - v^2/c^2)}. \tag{62}$$

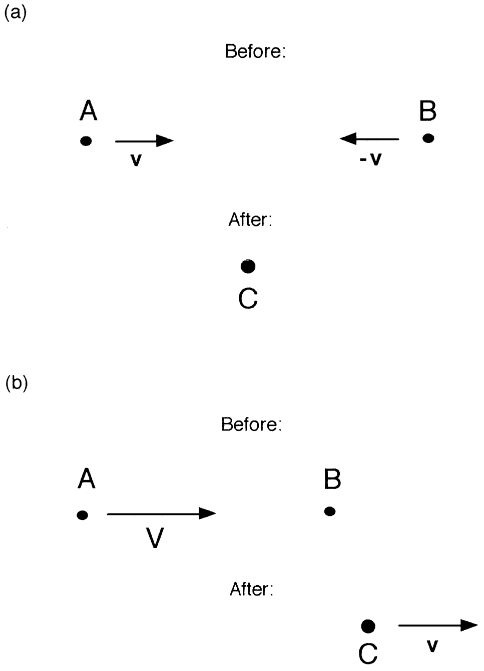


FIGURE 10 Inelastic collision of particles A and B. (a) As viewed in center of mass frame (chosen to be primed frame). (b) As viewed in rest frame of particle B (unprimed frame).

Object B is at rest in the unprimed frame, so the total incoming energy as observed in this frame is

$$E_{in} = E_A + E_B = mc^2\frac{(1 + v^2/c^2)}{(1 - v^2/c^2)} + mc^2 = \frac{2mc^2}{1 - v^2/c^2}. \tag{63}$$

After the collision, the composite object C (formed when A and B stick together) is at rest in the primed frame and moves with speed v in the unprimed frame. Thus the energy of C in the unprimed frame (which is the total energy going out from the collision as observed in this frame) is related to the total outgoing energy in the primed frame by

$$E_{out} = \frac{E'_{out}}{\sqrt{1 - v^2/c^2}}. \tag{64}$$

Applying momentum conservation in the unprimed frame and denoting the mass of C by M, we have

$$\frac{Mv}{\sqrt{1 - v^2/c^2}} = \frac{mV}{\sqrt{1 - V^2/c^2}} + m \cdot 0. \tag{65}$$

Substituting for V and $1 - V^2/c^2$ from Eqs. (60) and (61) we find

$$M = \frac{2m}{\sqrt{1 - v^2/c^2}}. \tag{66}$$

Clearly mass is not conserved in the collision, since the total mass of the incoming objects was $2m$. The final mass is increased—we will discuss this further. For the moment we simply note that unlike in Newtonian physics, *in relativity, the total mass is not a conserved quantity*.

The outgoing energy in the unprimed frame is $E_{out} = Mc^2/\sqrt{1 - v^2/c^2}$. From Eq. (66), this may be written

$$E_{out} = \frac{2mc^2}{1 - v^2/c^2}. \tag{67}$$

Comparing Eqs. (64) and (67) for E_{out} we see that the energy E'_{out} of C in the primed frame is given by

$$E'_{out} = \frac{2mc^2}{\sqrt{1 - v^2/c^2}}. \tag{68}$$

Thus, at least for this example and in the primed frame, we have shown that energy is conserved in the collision, since the energy of each of the incoming objects is $mc^2/\sqrt{1 - v^2/c^2}$. Furthermore, from Eqs. (63) and (67), it is obvious that conservation of energy holds in the unprimed frame as well.

C. Summary

1. Definitions of Momentum and Energy

In summary the relativistic definitions of momentum and energy of an object with mass m, in a frame where it is moving with velocity $\mathbf{v}$, are as follows:

$$\mathbf{p} = \gamma m\mathbf{v} = \gamma m\, d\mathbf{x}/dt = m\, d\mathbf{x}/d\tau. \tag{69}$$

$$\mathbf{E} = \gamma mc^2. \tag{70}$$

Here γ is the Lorentz factor $1/\sqrt{1 - v^2/c^2}$.

Sometimes the quantity γm is called the *relativistic mass* and m itself called the *rest mass*. Modern convention is to simply refer to m as the mass and to avoid the term "relativistic mass."

Energy and momentum are conserved in an isolated system, that is, in a given frame the value of the total energy (summed over all parts of the system) and the value of the total momentum does not change with time. Mass, however, is *not* conserved in general.

On the other hand, the mass of an object at any given moment is an invariant, that is, it is the same as observed in all frames. The values of the energy and momentum, however, depend on the observer's reference frame.

2. Relativistic Relationship Between Energy, Momentum, and Mass

Equations (57) and (58) are two correct interpretations of Einstein's famous equation $E = mc^2$. The general equation, however, is

$$E^2 = p^2c^2 + m^2c^4. \tag{71}$$

This is simply a rewrite of Eq. (56) in terms of $E = cp_t$ and $p = \sqrt{p_x^2 + p_y^2 + p_z^2}$. Eq. (71) is the correct relativistic relationship between the total energy, momentum, and mass of an object.

3. Lorentz Transformation of Momentum-Energy Four-Vector

We have shown that the component $p_t = E/c$ which we introduced to make a four-vector is just the total energy apart from the factor of c. Our discussion has shown that the *momentum-energy four-vector*, defined as $(E/c, p_x, p_y, p_z)$ satisfies a relation reminiscent of the invariance of the spacetime interval, i.e., the same combination of its components is also an invariant.

Like the space–time components t, x, y, z, the components of the momentum-energy four-vector transform according to the Lorentz transformation. Assuming that the primed frame moves with velocity v along the x-axis with respect to the unprimed frame, we have:

$$\begin{aligned} p'_x &= \gamma(p_x - \beta E/c) \\ p'_y &= p_y \\ p'_z &= p_z \\ E'/c &= \gamma(E/c - \beta p_x). \end{aligned} \tag{72}$$

4. Particles with Zero Mass

We noted that an object with nonzero mass can never reach or exceed the speed of light. However, there are particles with zero mass that travel with speed exactly c. The general relationship Eq. (71) between mass, energy, and momentum says that if $m = 0$ then

$$E = pc. \tag{73}$$

The quantum theory asserts that light has a dual wave-particle nature; the particles which make up light are particles with zero mass called photons. The relationship $E = pc$ is indeed consistent with all experiments and observations involving photons, for example, observations of collisions of photons with electrons (the Compton effect).

5. Conversion Between Different Forms of Energy

Note that even in nonrelativistic physics, if energy is to be conserved in an inelastic collision, then the kinetic energy of the incoming objects must go into some other form of energy—after all, in the center of mass frame, the kinetic energy is zero after the collision. For example, the initial kinetic energy of the incoming particles may be absorbed as an increase in the internal energy of motion of the molecules comprising the composite body formed in the collision (and perhaps its environment). This means that the temperature is very slightly increased. What is different in relativity is that when the internal energy and temperature change, so does the mass. The change in mass is extremely tiny in situations we are familiar with in everyday life, since the conversion factor c^2 between "mass" and "energy" is so huge. As we have already noted, mass (and thus rest energy) is *not* a conserved quantity in relativity.

Conversion of rest energy to other forms of energy occurs on a significant scale in nuclear reactions. These are of two basic types, fission reactions and fusion reactions. In a *fission* reaction, a single atomic nucleus dissociates into two or more pieces, where the sum of the masses of the pieces is less than the mass of the original nucleus. The difference in rest energy is transformed into kinetic energy of the pieces. In a fusion reaction, two nuclei fuse to form a nucleus that has less mass than the sum of original nuclei. The corresponding leftover rest energy is released partly as high energy electromagnetic radiation (including ordinary visible light). This is the mechanism by which the sun and other stars shine.

On a much smaller scale, the same type of thing happens when energy is released or absorbed in normal chemical reactions. The amount of energy involved is so small however, that the fractional difference in mass between the initial and final constituents is extremely tiny.

D. Force and Newton's Second Law in Relativity

Newton's Second Law, which relates the change in momentum of an object to the force **F** acting upon it, may be written in the form

$$\mathbf{F} = d\mathbf{p}/dt. \tag{74}$$

It turns out to be useful to define force such that this equation may still be used in relativity. In non-relativistic physics, Eq. (74) is equivalent to both $\mathbf{F} = m\mathbf{a}$ and $\mathbf{F} = m\mathbf{v}/dt$. This is not the case in relativity since $\mathbf{p} = \gamma m\mathbf{v}$ and γ depends on the velocity. In fact, in relativity the force and the acceleration are not necessarily even in the same direction.

As v approaches c, the Lorentz factor γ approaches infinity and thus so do the momentum and the energy. This is consistent with the observation that an object with nonzero mass can never quite reach the speed of light, no matter how much force is exerted on it.

VI. SPECIAL RELATIVITY AND ELECTROMAGNETISM

Maxwell's equations are one example of the postulate that fundamental laws of physics should be invariant with respect to changes from one inertial frame to another via a Lorentz transformation. Relativity shows that electric and magnetic fields are aspects of the same entity—the electromagnetic field. Indeed Einstein said,

> What led me more or less directly to the special theory of relativity was the conviction that the electromotive force acting on a body in motion in a magnetic field was nothing else but an electric field.

More precisely, when a pure electric field in one frame (e.g., the unprimed frame) is viewed from another frame (the primed frame) in motion with respect to the first, there can be a nonzero magnetic field in the second frame.

As usual we will assume the primed frame is moving with speed v along the x-axis, and that the x and x'axes are in the same direction. In this case, the transformation of the components of the electric field and the magnetic field is given by

$$\begin{aligned} E'_x &= E_x, & B'_x &= B_x \\ E'_y &= \gamma[E_y - (v/c)B_z], & B'_y &= \gamma[B_y + (v/c)E_z] \\ E'_z &= \gamma[E_z + (v/c)B_y], & B'_z &= \gamma[B_z - (v/c)E_y]. \end{aligned} \tag{75}$$

We shall not derive this result here. However, it follows from the Lorentz transformation plus one additional assumption.

This additional ingredient is Coulomb's Law, which was formulated in the 18th century. Coulomb's Law gives the force F exerted by one charge (with charge q_1) upon another charge (with charge q_2), assuming both charges are at rest:

$$F = k\frac{q_1 q_2}{r^2}. \tag{76}$$

Here r is the distance between the two charges, k is a constant, and the force F acts along the line through the two charges. If q_1 and q_2 have the same sign, the force is repulsive; if they have the opposite sign, the force is attractive. Coulomb's Law may be rewritten as

$$F = q_2 E, \tag{77}$$

where E is the magnitude of the electric field at q_2:

$$E = k\frac{q_1}{r^2}. \tag{78}$$

The magnetic field is an effect that may arise simply because one has changed reference frames. This is apparent from the above transformation. For example, suppose that in the unprimed frame there is no magnetic field, that is,

$B_x = B_y = B_z = 0$. From Eq. (75) we see that in the primed frame there are nonzero magnetic field components B'_y and B'_z transverse to the direction of motion of the primed frame if E_y or E_z is non-zero. Furthermore the transverse components of the electric field are different in the primed and unprimed frames.

Consider, for example a single point charge at rest in the unprimed frame. From Coulomb's law, we know that the electric field is radially symmetric as shown on the left hand side of Fig. 11. The field is stronger where the field lines are closer together. Suppose the charge is in motion with velocity v as shown on the right hand side of Fig. 11. Then the electric field is "flattened" along the direction of motion as shown. The field strength is increased transverse to the direction of motion and decreased parallel to the the direction of motion. In this particular example, $v/c = \sqrt{8/9}$, so that $\gamma = 3$. For larger v (and thus larger γ), the field would be flattened even more.

This transformation of the electric and magnetic fields is relevant in high energy particle accelerators, such as the linac discussed earlier. A linac accelerates a large number of charged particles, for example electrons, in a single short bunch. The mutual electrical repulsion of the electrons would be extremely strong if such a bunch were so short when at rest. However when the speed of the bunch is high enough, the forces between the electrons become almost negligible. This is due to two effects, both arising from the Lorentz transformation of the fields. One effect is that the electric field is mostly transverse as shown, so that it becomes small for electrons having different x coordinates. The other effect is that even for electrons having essentially the same x coordinate, the net force approaches zero as v approaches c, because the electric and magnetic forces tend to cancel each other in this limit. The calculation of the magnetic force on a charge is slightly more complicated than the calculation of the electric force—the magnetic force is perpendicular both to the direction of B and the direction of the velocity of the charge, and it depends on the magnitude of the velocity. However, the net result is that the electric and magnetic forces between two electrons having essentially the same x coordinate nearly cancel, if they are moving at sufficiently high speed along the x-axis.

Another way to see that the electrons are not much affected by each others' electric and magnetic fields is to take account of Lorentz contraction of the bunch. In the frame where the compact bunch is moving at very high speed, it is Lorentz contracted by a factor γ compared to its proper length. In the rest frame of the bunch, the bunch length (its proper length) is much longer so that the forces (given by Coulomb's Law) are reduced due to the increased distance between particles.

VII. SPECIAL RELATIVITY AND QUANTUM MECHANICS

Other developments occurring in parallel to Einstein's formulation of special relativity eventually led to an entirely new picture of the behavior of matter at the subatomic level, the realm of quantum phenomena. Quantum mechanics, a branch of physics that was motivated by studies of atomic structure and radiating bodies, eventually led to modifications of mechanics, in addition to those required

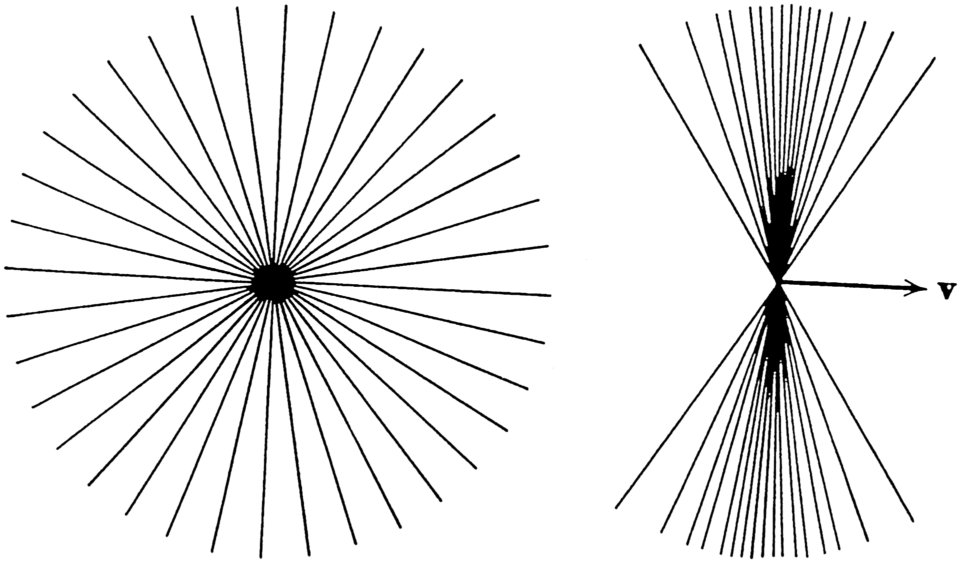


FIGURE 11 Electric field of a charged particle at rest (left), and of a charged particle moving with uniform velocity **v** such that $\gamma = 3$ (right). [Reproduced with permission from Jackson, J. D. (1975). "Classical electrodynamics," 2nd ed., p. 555, Wiley, New York.]

in going from Newtonian mechanics to special relativity. According to quantum mechanics, light exhibits both particle properties and wave properties, and so do other "objects" (e.g., electrons).

Special relativity retains in common with Newtonian mechanics the assumptions that (1) light is a wave, and (2) Newtonian determinism is valid (that is, there is no limitation in principle on how precisely we can simultaneously measure positions and velocities in a system, and we can use these measurements to predict the state of the system at later times). Quantum mechanics, on the other hand, says that there is an inherent uncertainty in the measurement process. The minimum uncertainty in the product of the position uncertainty Δx and the momentum uncertainty Δp is given by the Heisenberg uncertainty principle

$$\Delta x \, \Delta p \geq \hbar, \tag{79}$$

where $\hbar$ is Planck's constant. Planck's constant is a number so small that this uncertainty is irrelevant in everyday life where the objects we deal with are much larger than atoms.

The combination of special relativity and quantum mechanics has been very fruitful, leading to the formulation of quantum field theory. This synthesis was achieved by the middle of the 20th century, and is our current framework for describing and understanding the behavior of the most elementary particles observed in nature. As we have seen, the main motivation driving the development of special relativity was the desire to account correctly for electromagnetic phenomena. When electrodynamics was extended into the quantum domain, the result was quantum electrodynamics, the prototype quantum field theory. The underlying principles were further generalized to allow the incorporation of additional phenomena—in addition to electromagnetism, there exist other interactions between certain particles. These include the "strong" and "weak" forces, that are important, for example, in understanding how the nuclei of atoms are constructed.

A. Antiparticles

One consequence of quantum field theory is the necessity for the existence of *antiparticles*. This comes about because it is possible for Lorentz transformations to reverse the time order of two events x_1 and x_2 that have a space-like separation. "Space-like separation" means that the spatial separation of the two events is greater than c times their time separation. Thus, special relativity would say that there is no way for a particle to propagate from one event to the other, if quantum mechanics is ignored. But due to the uncertainty principle in quantum mechanics, there is a nonzero probability for a particle originating at x_1 to be absorbed at x_2 even when the two events have space-like separation. If the separation is space-like, in some other frame it can look like a particle originated at x_2 and was absorbed at x_1. If the particle was electrically charged, then from charge conservation its "time-reversed" form must appear to have the opposite charge. The mass is invariant, the same in all frames. Each type of charged subatomic particle is therefore required to have a corresponding type of antiparticle with opposite charge and identical mass.

A characteristic feature of very high energy collisions of subatomic particles is that new particle–antiparticle pairs are sometimes created. In some experiments, an electron may collide with its antiparticle (the positron) after both particles have been accelerated to very high energy. The electron and positron can annihilate each other—all of their rest energy plus their incoming kinetic energy is then available to produce other particles, so long as the sum of the rest energies of the produced particles is less than the original total energy and all other conservation laws (momentum, charge, etc.) are satisfied.

VIII. GENERAL RELATIVITY

The term "special" in special relativity refers to the fact that it is really only a restricted version of a more general theory of relativity which was put forth by Einstein in 1916. General relativity treats all reference frames—not just all inertial reference frames—on an equal basis. General relativity can readily be used to analyze situations in which observers are accelerating and/or gravitational fields are present.

SEE ALSO THE FOLLOWING ARTICLES

Cosmology • Electrodynamics, Quantum • Electromagnetics • Mechanics, Classical • Optical Interferometry • Particle Physics, Elementary • Quantum Theory • Relativity, General • Stellar Structure and Evolution • Time and Frequency

BIBLIOGRAPHY

Feynman, R. P., Leighton, R. B., and Sands, M. (1963). "The Feynman Lectures on Physics," Addison-Wesley, Reading, Massachusetts.

French, A. P. (1968). "Special Relativity," W. W. Norton and Company, New York.

Jackson, J. D. (1975). "Classical Electrodynamics," 2nd ed., Chapters 11 and 12, Wiley, New York.

Rindler, W. (1969). "Essential Relativity," Van Nostrand Reinhold, New York.

Sartori, L. (1996). "Understanding Relativity," University of California Press, Berkeley and Los Angeles, CA.

Stachl, J. (1998). "Einstein's Miraculous Year," Princeton University Press, Princeton, NJ.

Taylor, E. F., and Wheeler, J. A. (1992). "Spacetime Physics: Introduction to Special Relativity," 2nd ed., W. H. Freeman, New York.